numero 178
giugno 1983
anno XVI
volume XXX

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Una zona denuclearizzata in Europa

La proposta di mettere al bando le armi nucleari in una fascia dell'Europa centrale ai lati del confine Est-Ovest è una via politicamente possibile per ridurre il rischio di guerra nucleare

di Barry M. Blechman e Mark R. Moore

I colloqui sul controllo degli armamenti nucleari attualmente in corso a Ginevra fra gli Stati Uniti e l'Unione Sovietica hanno di mira due obiettivi: ridurre il numero delle armi nucleari strategiche, o a lungo raggio, spiegate dalle due superpotenze, e ridurre il numero delle armi nucleari a raggio intermedio, capaci di raggiungere un bersaglio in Europa. Dai negoziati è esclusa un'altra categoria molto importante, cioè quella delle armi nucleari tattiche, o a breve raggio. Delle circa 6000 armi nucleari attualmente installate in Europa dalle forze militari della NATO, più della metà sono classificate come tattiche. Quanto alle forze opposte del Patto di Varsavia, si pensa che abbiano schierato dalla loro parte del confine Est-Ovest, nell'Europa centrale, un arsenale analogo di armi nucleari. L'installazione in Europa delle armi nucleari a breve raggio viene intesa dagli uni e dagli altri come una possibilità di offrire un appoggio diretto alle forze di terra impegnate in combattimento: per questa ragione sono note anche col nome di «armi nucleari da campo». In questa categoria rientrano le testate nucleari per i missili balistici a breve raggio (intesi in questa sede come quei missili che hanno un raggio d'azione massimo di 150 chilometri), i proiettili d'artiglieria (con una gittata massima di 30 chilometri), i missili antiaereo e le mine esplosive.

Il numero considerevole di armi nucleari da campo con base in Europa presenta un problema del tutto particolare dal punto di vista del controllo degli armamenti. Essendo installate in gran numero, nell'Europa centrale, in posizioni avanzate vicino alla prima linea di qualsiasi eventuale guerra, esse hanno infatti buone probabilità di essere annientate dagli invasori. Dovendo scegliere fra perdere le armi o farne uso, la parte che si difende sarebbe spinta presumibilmente a impiegarle per prima, provocando in tal modo una *escalation* del conflitto fino a fargli varcare la «soglia nucleare».

Per questo, in quella acuta crisi politica che, con tutta probabilità, precederebbe lo scoppio delle ostilità potrebbero essere esercitate notevoli pressioni sui leader politici allo scopo di delegare in anticipo ai responsabili militari di stanza in Europa il potere di ricorrere, in determinate circostanze, alle armi nucleari da campo. C'è da chiedersi se un qualsiasi presidente americano o un qualsiasi leader europeo prenderebbe mai, in realtà, una decisione del genere; ciononostante, le pressioni per agire in questo senso sarebbero effettive, e non si può quindi escludere a priori questa possibilità. Anche escludendo però l'eventualità di una decisione deliberata di questo genere, il problema di mantenere, in tempo di guerra, l'effettivo controllo delle armi nucleari da campo sarebbe difficile. Considerata la confusione e l'incertezza delle informazioni che si associano inevitabilmente a operazioni militari su vasta scala, l'ubicazione specifica delle basi delle armi nucleari potrebbe essere un fattore fondamentale al momento di stabilire se prendere o no la decisione di dare il via a una guerra nucleare.

Un piano originale per affrontare questo problema fu presentato lo scorso anno dalla Independent Commission on Disarmament and Security Issues, un gruppo internazionale di attuali o ex funzionari governativi, costituitasi nel 1980 sotto la presidenza di Olof Palme, allora ex primo ministro svedese. (Oggi è di nuovo primo ministro.) Di questa commissione, i cui membri provenivano da nazioni dell'Est europeo, dell'Occidente e del Terzo Mondo, facevano parte fra gli altri Cyrus Vance, ex segretario di Stato degli Stati Uniti e Georgij Arbatov, membro del Comitato Centrale del partito comunista sovietico. In un suo rapporto, *Common Security: a Blueprint for Survival*, reso di pubblico dominio nel mese di giugno dello scorso anno, la commissione chiedeva l'istituzione in Europa di una «zona denuclearizzata relativamente alle armi da campo» unitamente ad altre «misure intese a rafforzare la soglia nucleare e a ridurre le pressioni per l'uso prematuro delle armi nucleari».

La zona denuclearizzata proposta potrebbe essere inizialmente - sempre secondo il rapporto - nell'Europa centrale, magari sotto forma di una fascia larga 150 chilometri ai due lati del confine tra Germania Occidentale e Germania Orientale e Cecoslovacchia. Col tempo essa verrebbe estesa dall'estremo settentrionale all'estremo meridionale dei due blocchi e vi sarebbero proibite le armi nucleari di qualsiasi genere.

Il concetto di zona denuclearizzata in Europa non è nuovo. Ciononostante, la proposta della commissione differisce da tutte le proposte precedenti per parecchi aspetti importanti, non ultimo dei quali è il fatto che essa si basa sulle caratteristiche di una particolare categoria di armi nucleari anziché su qualche altro criterio, come l'estensione territoriale delle nazioni interessate. A nostro parere il piano della commissione è un modo pratico e politicamente realizzabile per ridurre immediatamente il rischio di una guerra nucleare in Europa. In questo articolo, verranno compendiati gli argomenti a favore di una zona centro-europea denuclearizzata relativamente alle armi da campo. In particolare verranno spiegate le ragioni per le quali si è convinti che questo specifico piano potrebbe avere successo là dove tutti gli altri hanno fallito.

Quasi tutte le nazioni si sono già dichiarate d'accordo sul fatto che certi luoghi debbano rimanere denuclearizzati. L'Antartide, lo spazio esterno e i corpi celesti, i fondali marini e l'America Latina sono soggetti tutti a trattati che vietano nel loro ambito lo spiegamento di armi nucleari. Sono state avanzate proposte per proibire le armi nucleari anche nel Medio Oriente, nell'Asia meridionale, nel Pacifico meridionale e in Africa, nonché in Europa. Ognuna è stata oggetto di almeno qualche discussione in vari consessi internazionali. Ciononostante, questi sforzi per creare zone denuclearizzate hanno prodotto di gran lunga più proposte che risultati. Non solo, ma tali proposte hanno dato spesso l'impressione di non essere tanto tentativi sinceri per ridurre il rischio di una guerra nucleare quanto piuttosto manovre diplomatiche

La zona denuclearizzata nell'Europa centrale, che è stata proposta, non dovrebbe avere armi nucleari. Il rischio che, nel caso di un attacco militare convenzionale, la parte assalita sia indotta a lanciare per prima le proprie armi nucleari da campo per impedire che vengano sopraffatte dalle forze assalitrici sarebbe così ridotto. Secondo la commissione Palme, gli esatti limiti geografici della zona «dovrebbero essere definiti per mezzo di trattative» fra le parti interessate. A fini illustrativi, comunque, ha suggerito una zona formata all'inizio da una fascia larga 150 chilometri (*in colore intenso*) da un lato e dall'altro del confine della Germania Occidentale con la Germania Orientale e con la Cecoslovacchia. A differenza di precedenti proposte di questo tipo, in alcune parti dei tre paesi sarebbe sempre possibile installare armi nucleari (*in colore chiaro*).

NOTA REDAZIONALE

La Independent Commission on Disarmament and Security Issues, che nella sua relazione finale, resa di pubblico dominio nel giugno del 1982, invitava fra l'altro a istituire una zona denuclearizzata relativamente alle armi da campo, era composta da 18 membri, presenti tutti in veste privata. Il presidente della commissione era Olof Palme, a quel tempo membro del Parlamento svedese, presidente del partito socialdemocratico svedese ed ex primo ministro. (Oggi Palme è di nuovo primo ministro.) Gli altri componenti della commissione erano Georgij Arbatov, membro del Comitato Centrale del partito comunista sovietico, deputato del Soviet Supremo e direttore dell'Istituto per gli Stati Uniti e il Canada, una sezione dell'Accademia delle scienze; Egon Bahr, membro del Parlamento della Repubblica Federale Tedesca, presidente del sottocomitato per il disarmo e il controllo degli armamenti di quell'organismo ed ex ministro per la cooperazione economica; Gro Harlem Brundtland, membro del Parlamento norvegese, presidente del partito laburista norvegese ed ex primo ministro di Norvegia; Jozef Cyrankiewicz, ex primo ministro di Polonia ed ex presidente del Consiglio di Stato; Jean-Marie Daillet, membro del Parlamento francese, vice presidente del comitato difesa di quell'organismo e presidente del comitato difesa della UDF, il partito giscardiano; Robert A. D. Ford, consigliere speciale del governo canadese per le relazioni Est-Ovest, ex ambasciatore canadese in Colombia, Jugoslavia, Egitto e Unione Sovietica; Alfonso Garcia-Robles, ambasciatore e presidente della delegazione messicana al Committee on Disarmament delle Nazioni Unite ed ex ministro degli esteri del Messico (nel 1982, vincitore ex aequo del premio Nobel per la pace per la parte svolta nella creazione di una zona latino-americana denuclearizzata); Haruki Mori, ex ambasciatore giapponese in Inghilterra e presso l'OCSE ed ex vice ministro degli affari esteri; C. B. Muthamma, ambasciatore indiano presso i Paesi Bassi ed ex ambasciatore in Ghana e in Ungheria; Olusegun Obasanjo, generale delle forze armate nigeriane, membro del Consiglio di Stato, membro emerito dell'Università di Ibadan ed ex capo di Stato della Nigeria; David Owen, membro del Parlamento britannico ed ex segretario di Stato per gli affari esteri e per gli affari del Commonwealth; Shridath Ramphal, segretario generale del Commonwealth ed ex ministro degli esteri della Guyana; Salim Salim, ministro degli affari esteri della Tanzania; Soedjatmoko, rettore della United Nations University di Tokio ed ex ambasciatore indonesiano presso gli Stati Uniti; Joop den Uyl, membro del Parlamento e vice primo ministro dei Paesi Bassi, leader del partito laburista olandese ed ex primo ministro; Cyrus Vance, ex segretario di Stato degli Stati Uniti.

	ARMA E ANNO DI SPIEGAMENTO		RAGGIO D'AZIONE (CHILOMETRI)	POTENZA ESPLOSIVA MASSIMA (CHILOTON)	DISPOSITIVI DI LANCIO
NATO	LANCE	1973	115	50	90
	HONEST JOHN	1953	38	20	90
	CANNONE M 109 DA 155 mm	1964	18	2	1455
	CANNONE M 110 DA 203 mm	1962	16	1	390
	PLUTON (FRANCIA)	1974	120	10	42
PATTO DI VARSAVIA	FROG/SS-21	1950/1977	75/155	200	375
	SCUD-A	1957	85	1	250
	CANNONE DA 203 mm	1979	31	?	150
	CANNONE DA 240 mm	1979	30+	?	150

Le armi nucleari da campo attualmente spiegate nell'Europa centrale, da un lato e dall'altro del confine Est-Ovest, comprendono i missili balistici a breve raggio d'azione e i pezzi d'artiglieria elencati in questa tabella. Per le forze della NATO sono inclusi tutti quei sistemi destinati al teatro delle operazioni europeo. (La Francia, pur essendo membro della NATO, non fa parte a pieno diritto dell'alleanza militare.) Per le forze del Patto di Varsavia, sono compresi i sistemi spiegati al di fuori del territorio dell'Unione Sovietica e nei suoi distretti militari occidentali. Tutti i missili indicati sono capaci di portare sul bersaglio soltanto una testata nucleare ciascuno. Non figurano le mine esplosive nucleari e i missili superficie-aria con testata nucleare, che a volte vengono fatti rientrare nel gruppo delle armi nucleari tattiche. Sono escluse anche le migliaia di armi nucleari a raggio più lungo, spiegate sulla terraferma o in mare alla portata di bersagli europei.

per ottenere vantaggi di natura politica.

Il primo e più particolareggiato piano per una zona denuclearizzata in Europa fu presentato nel 1957 da Adam Rapacki, a quel tempo ministro degli esteri polacco. Il piano Rapacki (come fu chiamato in seguito) proponeva che i governi della Polonia, della Cecoslovacchia, della Repubblica Democratica Tedesca e della Repubblica Federale Tedesca decidessero di comune accordo di non fabbricare, mantenere o far installare armi nucleari entro i propri confini. La denuclearizzazione della zona sarebbe stata garantita dai paesi già dotati allora di armi nucleari (Stati Uniti, Unione Sovietica e Regno Unito) e, per assicurarne l'osservanza, sarebbe stato istituito un sistema di sorveglianza aerea e terrestre. Nella sua formulazione originaria, il piano Rapacki non prevedeva un trattato formale per l'istituzione della zona in questione, ma si basava su dichiarazioni unilaterali dei governi interessati.

Nella fotografia un missile americano Lance viene approntato per un lancio di prova da un veicolo lanciamissili a propulsione autonoma. Il razzo a combustibile liquido può portare una testata o nucleare o convenzionale. Il suo raggio d'azione massimo è di circa 115 chilometri. Per gli Stati Uniti, il Lance ha sostituito l'Honest John quale missile nucleare standard da campo. Altre nazioni che fanno parte dell'alleanza atlantica hanno nel proprio arsenale tanto il Lance quanto l'Honest John. Tutte le testate nucleari per i missili sono sotto la sorveglianza degli Stati Uniti.

Gli obiettivi politici del piano Rapacki furono subito evidenti alle nazioni occidentali. Il piano, infatti, veniva proposto in un momento in cui i paesi della NATO stavano discutendo attivamente sulla possibilità di controbilanciare per mezzo di uno «scudo nucleare» la superiorità delle forze militari convenzionali (non nucleari) del Patto di Varsavia. La decisione di seguire questa linea di condotta fu presa in effetti poco dopo, nel dicembre del 1957, a una riunione del North Atlantic Council. Di conseguenza, i paesi della NATO considerarono il piano Rapacki un tentativo per indebolire la loro posizione militare nell'Europa centrale, e lo respinsero.

Il piano polacco fu modificato e ripresentato alle nazioni occidentali due volte, nel novembre del 1958 e nel marzo del 1962. Le modifiche non rispondevano però alle obiezioni di fondo della NATO e, quindi, non fecero molta strada. Successive proposte di zone denuclearizzate in Europa, avanzate dai paesi dell'Est, si sono concentrate solo su particolari subregioni. I Balcani e l'area dell'Adriatico sono da lungo tempo oggetto di proposte, così come lo è l'intero bacino del Mediterraneo. In anni recenti, la regione nordica, che il defunto premier sovietico Nikita Kruscev, nel 1959, fu il primo a indicare come zona adatta per essere denuclearizzata, ha ricevuto molta attenzione da parte dei funzionari sovietici.

Non molto tempo dopo che erano state avanzate le prime proposte dei paesi dell'Est per la creazione di zone denuclearizzate, gli Stati Uniti elencarono quattro criteri che un'iniziativa del genere avrebbe dovuto soddisfare per ottenere l'appoggio americano: (1) la proposta avrebbe dovuto essere promossa spontaneamente dagli stati della zona interessata; (2) la proposta avrebbe dovuto comprendere tutti gli stati della zona interessata, o almeno tutti quelli dotati di una grande potenza militare; (3) non avrebbero dovuto essere turbati gli accordi esistenti per la sicurezza e per l'equilibrio militare; (4) l'accordo avrebbe dovuto comprendere alcune clausole per la verifica dell'ottemperanza.

Nel caso dell'America Latina, gli Stati Uniti trovarono evidentemente che la proposta di creare una zona denuclearizzata si adattava a questi criteri, o poteva almeno essere modificata in tal senso. D'altra parte, gli Stati Uniti e altri governi occidentali non hanno reputato che precedenti proposte per la creazione di zone denuclearizzate in Europa potessero essere oggetto di trattative. In alcuni casi le obiezioni che hanno avanzato erano di carattere tecnico: considerata l'opposizione posta da lunga data da parte dell'Unione Sovietica a ispezioni in loco, il com-

Il cannone statunitense M 110 da 203 millimetri, fornito dagli Stati Uniti a quasi tutti i paesi della NATO, in un tiro di prova con un proiettile non nucleare. L'obice da otto pollici è anche in grado di sparare proiettili nucleari. La sua gittata massima è di circa 16 chilometri; può essere montato o su un veicolo a propulsione autonoma, come nella fotografia, o su un carrello trainato. La struttura posteriore assorbe il rinculo.

pito di assicurare il ritiro e la continuata assenza di armi nucleari da una zona che in precedenza ne abbondava è sembrato difficile, se non impossibile.

Le obiezioni fondamentali dell'Occidente a zone denuclearizzate in Europa sono derivate però da altri due fattori: l'asimmetria di fondo nelle caratteristiche geopolitiche degli Stati Uniti e dell'Unione Sovietica e il ruolo fondamentale che le armi nucleari svolgono ormai nella strategia militare occidentale e nelle relazioni politiche fra i paesi membri della NATO. Dato il peso di queste considerazioni, i funzionari occidentali hanno tendenzialmente considerato le proposte dei paesi dell'Est, per la creazione di zone denuclearizzate in varie parti d'Europa, per lo più propagandistiche, volte a incoraggiare il dissenso fra le nazioni dell'Europa occidentale e ad allentare i legami politici fra i paesi membri dell'alleanza atlantica. È bene valutare appieno queste riserve della NATO se si vogliono ideare misure realistiche per ridurre il rischio di una guerra nucleare in Europa.

Lungo tutta la storia dell'alleanza occidentale, i funzionari pubblici e i comuni cittadini delle nazioni occidentali hanno discusso fino a che punto sarebbe efficace e prudente basarsi sulle armi nucleari per conservare la pace in Europa. I fautori dei sistemi di difesa convenzionali sostengono che la via più sicura per la stabilità europea è quella di mantenere abbastanza forte il contingente militare convenzionale così da persuadere i leader dell'Unione Sovietica che, nel caso di un conflitto, essi sarebbero sconfitti sul campo di battaglia e si troverebbero quindi a dover scegliere fra la sconfitta o l'escalation della guerra fino al punto di ricorrere all'uso delle armi nucleari. Si sostiene che i leader sovietici, se si convincessero che il probabile risultato di un conflitto in Europa sarebbe una scelta di questo genere, deciderebbero contro la guerra, favorendo invece relazioni stabili fra le nazioni dell'Est e quelle dell'Occidente. Il sostegno dato a questo modo di vedere si è manifestato con ricorrenti tentativi di rafforzare le forze convenzionali della NATO.

Per molte altre persone in Occidente, la minaccia di una guerra nucleare e le forze necessarie per rendere credibile tale minaccia sono sempre state, invece, l'essenza della strategia della NATO. Da questo punto di vista, ciò che persuade i leader dell'Unione Sovietica a cercare la stabilità e a evitare la guerra in Europa non è tanto la previsione che potrebbero perdere un conflitto di questo genere, quanto piuttosto un'intuizione generale che, se gli eventi dovessero sfuggire di mano, i risultati di qualsiasi impegno militare potrebbero essere disastrosi. In realtà, coloro i quali la pensano in questo modo sostengono che le forze militari della NATO dovrebbero essere intese a evitare una situazione in cui i risultati di un conflitto diventino, a un certo punto, prevedibili. Secondo costoro, se è possibile prevedere il risultato, non è da escludere che, in certe circostanze, esso diventi accettabile. A loro dire, sarebbe molto meglio aumentare il potere di dissuasione, inculcando con forza l'idea che il risultato del conflitto sarebbe ineluttabilmente la catastrofe.

Chi ragiona così cerca di far capire ai leader dell'Unione Sovietica che qualsiasi atto aggressivo manifesto in Europa potrebbe facilmente deteriorarsi in un conflitto, portando via via a una escalation verso una guerra nucleare totale fra Stati Uniti e Unione Sovietica. In breve, cerca di sostenere una posizione in cui il pericolo estremo di un confronto in Europa - l'olocausto nucleare - sia il più possibile bloccato. Nessun miglioramento nelle forze convenzionali può servire a questo scopo. Da questo punto di vista, anzi, quanto più le difese convenzionali appaiono forti, tanto maggiore è il rischio che i leader dell'Unione Sovietica sbaglino i calcoli e decidano per un'aggressione, nella convinzione che tali difese abbiano ridotto il rischio di una guerra nucleare.

Anche per questa ragione la NATO non ha mai tenuto fede al dichiarato obiettivo di rafforzare le proprie difese convenzionali ed è sempre per questo motivo che dichiara esplicitamente che inizierebbe e potenzierebbe una guerra nucleare in Europa, se necessario. Sempre per questa ragione, infine, mantiene nell'Europa occidentale migliaia di armi nucleari tattiche e sta spiegando una nuova generazione di armi nucleari a raggio intermedio, esprimendo regolarmente le proprie preoccupazioni a proposito dell'equilibrio tra forze nucleari strategiche statunitensi e sovietiche.

Alla base di tutta l'importanza che viene attribuita alla difesa nucleare, c'è però ancora un altro motivo. Da quasi quarant'anni, gli strateghi e i funzionari politici americani sono convinti che la peggiore minaccia agli interessi degli Stati Uniti è rappresentata dall'Unione Sovietica. Si ritiene che l'Unione Sovietica non solo minacci la sicurezza degli Stati Uniti in termini militari, ma anche metta in di-

scussione, in modo più indefinibile, i valori della società americana. Di conseguenza, quando la strategia americana viene considerata in termini geopolitici, si nota che essa ha mirato essenzialmente, in tutto questo periodo, a contenere sia il potere militare sia l'influenza politica dell'Unione Sovietica nel continente euroasiatico. È della massima importanza il fatto che la strategia statunitense abbia cercato di limitare l'influenza dell'Unione Sovietica in quelle parti dell'Europa che si trovano al di là della fascia di nazioni occupate dalle truppe sovietiche nel 1945. Il portavoce di elezione di questa strategia è, ovviamente, la NATO.

Nella lotta fra le due superpotenze per l'influenza sull'Europa occidentale, gli Stati Uniti hanno un netto svantaggio geografico. La potenza militare dell'Unione Sovietica è una presenza costante, fortemente avvertita nelle capitali dell'Europa occidentale. Per contro, indipendentemente dall'enorme interesse economico e politico degli Stati Uniti perché l'indipendenza e l'orientamento filoamericano dell'Europa occidentale siano mantenuti, l'impegno statunitense per la sicurezza e per la presenza militare in Europa viene oggi avvertito, e sempre lo sarà, come qualcosa di transitorio e di non sicuro affidamento. Le nazioni dell'Europa occidentale si rendono conto che gli Stati Uniti hanno sempre la possibilità di ritirarsi al di là di quell'oceano di cui si sono fatti scudo in passato. Questi timori subiscono spesso un aggravamento, nei momenti in cui gli Stati Uniti mostrano di avere uno dei loro ricorrenti accessi di neoisolazionismo.

Per controbilanciare l'asimmetria che esiste, tanto come percezione quanto come realtà, nella presenza militare sovietica e americana nell'Europa occidentale, i detentori del potere decisionale negli Stati Uniti hanno fatto continui sforzi per stringere sempre più i legami tra forze di stanza nell'Europa occidentale e potenza militare globale statunitense e per rassicurare gli alleati (e ammonire gli avversari) sul carattere duraturo degli impegni presi dagli Stati Uniti. Questi sforzi assumono varie forme, ma le componenti di fondo sono state le iniziative intese ad «associare» le forze nucleari strategiche americane alle forze di terra spiegate in Europa.

La strategia occidentale in Europa si basa sul concetto di una escalation di risposte all'aggressione, cominciando da contrattacchi con armi convenzionali nel caso in cui l'attacco da parte dei paesi del Patto di Varsavia si limiti alle forze convenzionali, ma non escludendo a priori la possibilità di avviare e continuare l'escalation verso una guerra nucleare. Questo atteggiamento dichiarato viene avvalorato dallo spiegamento in Europa di armi nucleari a breve raggio d'azione non lontano dalle presunte prime linee; in molti scenari queste armi, di cui è previsto l'uso contro forze militari sul campo di battaglia, sarebbero le prime a essere impiegate se si arrivasse alla decisione di varcare la soglia nucleare. Poiché hanno una potenza esplosiva relativamente ridotta e sarebbero impiegate contro obiettivi militari, si pensa che possano rendere più credibile la minaccia di dare il via a una guerra nucleare.

Per dare ulteriore credibilità alla minaccia nucleare della NATO, l'alleanza atlantica ha intenzione di installare nell'Europa occidentale missili americani a medio raggio in grado di colpire bersagli in territorio sovietico. Queste armi a medio raggio d'azione potrebbero essere lanciate, sempre contro obiettivi militari, se l'uso delle armi nucleari da campo di battaglia non fosse sufficiente a porre termine al conflitto. La mossa sarebbe presumibilmente meno difficile di quella successiva, cioè del ricorso alle forze strategiche principali degli Stati Uniti. Anche così, tuttavia, gli attacchi nucleari su territorio sovietico da parte di missili controllati dagli Stati Uniti porterebbero quasi certamente a una risposta sovietica contro gli Stati Uniti facendo precipitare i tempi di un coinvolgimento delle forze strategiche principali dell'una e dell'altra parte. L'idea è che i leader dell'Unione Sovietica, rendendosi perfettamente conto di questa escalation, siano dissuasi dal fare il primo passo. Sul piano teorico, la deterrenza si rafforza nella misura in cui viene reso credibile il coinvolgimento finale delle forze nucleari strategiche degli Stati Uniti.

Questa ripartizione del rischio è, in un certo senso, il legame fondamentale che, sul piano pratico e psicologico, unisce l'alleanza atlantica, almeno per quel che riguarda la politica di sicurezza. Promettendo di ricorrere, se necessario, alle forze nucleari strategiche in difesa dell'Europa occidentale, e accettando in tal modo l'eventualità di rappresaglie sovietiche sul territorio americano, gli Stati Uniti sono entrati simbolicamente come soci alla pari del destino dell'Europa, controbilanciando così il proprio isolamento geografico.

Dovrebbero essere lampanti, sulla base di questa argomentazione, le manchevolezze nelle proposte, fatte dai paesi dell'Est, di denuclearizzare alcune zone europee, così come vengono viste nelle capitali della NATO. Nella proposta tradizionale, una zona denuclearizzata priverebbe l'Europa occidentale di un legame fondamentale tra le proprie difese tradizionali e il deterrente nucleare strategico degli Stati Uniti, accentuando in tal modo le esistenti asimmetrie sia nella posizione geografica delle due parti sia nell'equilibrio delle forze militari convenzionali. Questa argomentazione vale soprattutto per le proposte (come il piano Rapacki) fatte per creare delle zone denuclearizzate nell'Europa centrale, una subregione che è al centro di tutte le discussioni sulla difesa dell'Europa occidentale. Proposte analoghe per altre subregioni condividono alcune delle stesse insufficienze politiche e militari e, in ogni caso, sono considerate inutili precedenti che potrebbero rendere più difficile il rifiuto di una zona denuclearizzata nell'Europa centrale.

Dal punto di vista della NATO, i punti deboli delle zone denuclearizzate proposte tradizionalmente sono, pertanto, i seguenti. In primo luogo, le iniziative non si accompagnano di norma a proposte intese a porre limiti alle forze convenzionali. Ciò significa che la capacità della NATO di controbilanciare con la minaccia di escalation nucleare la superiorità delle forze convenzionali del Patto di Varsavia verrebbe indebolita. È ovvio che la NATO potrebbe mobilitare le proprie maggiori risorse economiche per reggere in maniera più efficace il confronto in termini di forze militari convenzionali, ma la cosa verrebbe a costare molto e sembra poco realistica ove si consideri la tradizionale riluttanza delle democrazie occidentali a sostenere pesanti spese militari in tempo di pace.

Inoltre, anche se si sviluppassero difese convenzionali più efficaci, rimarrebbe sempre la questione di quanto l'attenuazione della minaccia nucleare della NATO ridurrebbe la capacità delle nazioni dell'Europa occidentale di resistere alle pressioni politiche e perfino all'aggressione militare dell'Unione Sovietica. Pur essendo impossibile da dimostrare, il fatto che le relazioni internazionali in Europa si siano evolute negli ultimi 35 anni in una configurazione abbastanza stabile fa pensare che la strategia della NATO possa benissimo aver adempiuto le aspettative militari e politiche dei suoi ideatori. Come minimo, esso indica che la dottrina esistente non dovrebbe essere modificata senza che vi siano fondate ragioni per ritenere che le alternative a essa abbiano un esito altrettanto positivo e siano più adatte alle circostanze attuali.

In secondo luogo, la rimozione delle forze nucleari dall'Europa centrale, poiché farebbe pensare a un distacco del deterrente nucleare strategico americano dalle difese dell'Europa occidentale, potrebbe far presagire una separazione politica dell'Europa occidentale dagli Stati Uniti. Si teme che una evoluzione di questo genere porterebbe inevitabilmente prima a un aumento dell'influenza e poi al predominio dell'Unione Sovietica nelle questioni europee. Il fatto che il territorio sovietico non fosse incluso nelle proposte di denuclearizzazione dell'Europa centrale dava particolare risalto al problema. In realtà, ciò significa che, se la NATO avesse accettato la proposta di Rapacki o altre proposte successive, avrebbe tacitamente acconsentito a una diminuzione della presenza americana nell'Europa occidentale senza nessuna adeguata diminuzione della presenza sovietica. Le implicazioni politiche di un magro affare di questo genere sono evidenti. Infatti esso sarebbe stato generalmente interpretato in modo da far pensare che all'Unione Sovietica venga concesso di avere un'influenza decisiva nelle questioni dell'Europa occidentale.

Un altro punto entra nelle valutazioni della NATO. Oltre al fine di contenere la potenza sovietica, l'associazione formale in un'alleanza militare della Germania Occidentale ad altri stati dell'Europa occidentale e dell'America Settentrionale aveva lo scopo di contribuire a sanare le ferite della seconda guerra mondiale. Insieme alle varie istituzioni economiche multinazionali create negli anni cinquanta, la NATO è riuscita a favorire la riconciliazione della Germania con i suoi ex nemici, rendendo possibile in Europa lo sviluppo di relazioni stabili e l'eliminazione delle tensioni nazionali che, per ben tre volte nell'arco di un secolo, avevano portato allo scoppio di guerre disastrose.

Il piano Rapacki avrebbe posto la Germania Orientale e quella Occidentale (insieme con la Polonia e la Cecoslovacchia) in una categoria speciale. La Germania Occidentale non sarebbe stata più un membro a pieno diritto dell'alleanza atlantica. Inoltre, una decisione del genere non sarebbe stata presa dagli stessi tedeschi, ma sarebbe stata loro imposta da forze esterne. Questo avrebbe accentuato la posizione speciale della Germania e si temeva che avrebbe rafforzato l'eventualità di una ricomparsa, all'interno della Germania stessa, di forze nazionaliste che avrebbero potuto mettere a repentaglio prima i legami politici fra la Germania e le altre nazioni occidentali e poi quelli fra nazioni dell'Est europeo e nazioni occidentali. In sintesi, isolando la Germania, il piano Rapacki e altre proposte di zone denuclearizzate nell'Europa centrale fecero nascere il timore che si ricreassero vecchie tensioni. Di gran lunga migliore, per molti leader occidentali, risultava la prospettiva dell'ininterrotta integrazione della Germania nel resto dell'Europa occidentale e del graduale miglioramento delle relazioni fra i due blocchi nel loro complesso.

Il piano Rapacki e le successive proposte di zone denuclearizzate nell'Europa centrale erano, dunque, gravate in partenza dal peso delle loro implicazioni politiche. Facendo intravedere una netta differenziazione di rischio fra Stati Uniti ed Europa, essi furono interpretati come un mezzo per erodere l'influenza americana in Europa e per accrescere il potere sovietico. Si ebbe la percezione, proprio perché accordavano alla Germania un trattamento particolare, che avrebbero probabilmente minato alle fondamenta tutti gli sforzi che erano stati fatti per integrare la Germania Occidentale nella comunità delle nazioni occidentali. L'attuale proposta di una zona denuclearizzata relativamente alle armi da campo, avanzata dall'International Commission on Disarmament and Security Issues, è stata studiata attentamente in modo proprio da evitare questi inconvenienti.

La Commissione ha tenuto in debita considerazione le complesse relazioni fra tipi diversi di forze militari in Europa e ha sollecitato un'impostazione organica del problema del controllo degli armamenti quale unico modo per assicurare il conseguimento di livelli inferiori di forze e la riduzione del rischio di una guerra. La commissione ha dichiarato nella sua relazione che «un accordo per raggiungere una certa parità tra le forze convenzionali dei due blocchi... faciliterebbe una riduzione delle armi nucleari e una risistemazione della priorità, oggi accordata agli armamenti nucleari nella pianificazione di contingenti militari».

Di conseguenza, la commissione ha sollecitato all'Europa la convocazione di una conferenza dei ministri degli esteri dei paesi della NATO e del Patto di Varsavia, che da 10 anni a questa parte stanno negoziando a Vienna un accordo sulle forze

Questi missili sovietici a breve raggio d'azione sono trasportati su terreno fangoso con veicoli cingolati; chiamati SCUD-A dalle fonti di controspionaggio occidentali, possono portare una testata o nucleare o convenzionale. La loro gittata massima è di circa 85 chilometri. Le testate nucleari per le armi del Patto di Varsavia sono sotto la sorveglianza dell'Unione Sovietica.

convenzionali. Essa ha invitato i ministri degli esteri a raggiungere un primo accordo che garantisca la parità fra le due parti e renda possibili trattative continue per ridurre in misura consistente le forze convenzionali. Secondo i membri della commissione, le differenze che rimangono fra le due parti potrebbero essere eliminate abbastanza rapidamente se ci fosse la volontà politica di farlo. Concludendo, la commissione afferma che sia l'impulso sia l'opportunità per portare avanti le cose potrebbero essere forniti da un accordo che preveda una riunione a livello dei ministri degli esteri.

La commissione ha prestato notevole attenzione anche alle armi chimiche, sollecitandone l'eliminazione da tutta l'Europa e proponendo misure associate per assicurare che entrambe le parti si attengano a tale veto. Queste misure comprenderebbero ispezioni in loco «su richiesta»; ciò significa che, una volta all'anno, a ognuna delle due parti è consentito, nel caso in cui essa voglia avvalersi di questa possibilità, di ispezionare un certo numero di attrezzature di propria scelta sul territorio della parte avversa. Si tratta di una proposta importante e originale. Il fatto che il rappresentante sovietico della commissione abbia accettato la clausola relativa all'ispezione fa pensare a una maggiore comprensione da parte sovietica, rispetto al passato, delle esigenze occidentali per una seria considerazione dei problemi relativi alla verifica degli accordi sul controllo degli armamenti. La sensibilità sovietica in merito alle preoccupazioni relative a tale verifica è emersa anche nelle discussioni sulla proposta di una zona denuclearizzata.

La commissione ha considerato parecchie soluzioni alternative per la componente nucleare dei problemi europei attinenti alla sicurezza, ivi compresa la creazione di una zona denuclearizzata definita da confini nazionali. Alla fine, però, i suoi membri hanno deciso di perseguire un'impostazione funzionale, concentrandosi su categorie specifiche di armi che pongono problemi particolari.

La commissione ha dedicato particolare attenzione ai «pericoli posti in Europa da quelle armi nucleari i cui vettori sono dislocati in numero notevole in posizioni avanzate». Essa ha sollecitato la reciproca rimozione di armi nucleari da campo di battaglia dalle zone vicine al confine Est-Ovest, risultato da conseguire attraverso un negoziato che istituisca una zona priva di tali armi. Per ridurre al minimo le difficoltà di una verifica, essa ha anche proposto che, nella zona in questione, non siano consentite munizioni nucleari di alcun genere. Essa ha altresì sottolineato l'opportunità di proibire anche i depositi per le munizioni nucleari, le manovre che simulano operazioni nucleari e i preparativi per la collocazione di mine nucleari. Infine, ha suggerito di negoziare norme speciali per lo spiegamento, nella zona, di artiglieria «a doppio uso» e di missili a breve raggio d'azione che possano essere usati per portare sul bersaglio munizioni nucleari.

La commissione ha ritenuto più opportuno non raccomandare confini specifici per la zona, lasciando questa decisione alle parti interessate. Per fini illustrativi, tuttavia, è stata citata la possibilità di istituire una zona di 300 chilometri, centrata sul confine Est-Ovest *(si veda l'illustrazione a pagina 17)*. Una fascia del genere includerebbe inizialmente parti della Germania Occidentale, della Germania Orientale e della Cecoslovacchia. Però, a differenza del piano Rapacki e di successive proposte fatte dai paesi dell'Est, parti di tutte e tre le suddette nazioni dovrebbero essere sempre disponibili per spiegamento di armi nucleari; è, questa, una distinzione di importanza fondamentale. A mano a mano che la zona venisse estesa sia a nord sia a sud, si potrebbero aggiungere a essa, a seconda del modo in cui fossero fissati i confini, parti dell'Austria, dell'Ungheria, della Jugoslavia, dell'Italia, della Danimarca, della Svezia, della Norvegia, della Finlandia e dell'Unione Sovietica.

La commissione ha anche lasciato ai negoziatori il compito di definire le norme specifiche, che sono necessarie per assicurare alle parti la possibilità di verificare l'ottemperanza alle norme dell'accordo. I membri hanno, tuttavia, concordato che le norme in questione «includessero un numero limitato di ispezioni in loco su richiesta». Anche in questo caso è significativo il fatto che il rappresentante sovietico abbia accettato di includere questa frase. Sebbene Arbatov, al pari di tutti gli altri membri della commissione, fosse presente in veste privata, è evidente che il suo assenso al testo - e in particolare alle raccomandazioni - non è stato dato senza autorizzazione superiore. L'accettazione dell'idea di ispezioni in loco in questo contesto è in armonia con la tendenza dei sovietici verso una maggiore ricettività nei riguardi di mezzi di verifica più invadenti.

La commissione ha raccomandato, infine, altre misure per rafforzare ulteriormente la soglia nucleare e ridurre le pressioni per un uso prematuro delle armi nucleari. Molto acuta è stata la raccomandazione fatta da una maggioranza di membri della commissione di non installare in Europa armi nucleari che potrebbero apparire di più facile impiego, soprattutto piccole armi nucleari come quelle a radiazione intensificata («bombe a neutroni»). Anche se non è stata fatta alcuna raccomandazione specifica, in una sezione della relazione si è anche discusso dei problemi speciali associati ai missili antiaereo con testate nucleari e alle mine esplosive, facendo pensare che l'eliminazione di queste armi dall'Europa potrebbe rafforzare ulteriormente le barriere che si frappongono all'avvio di una guerra nucleare.

In breve, sembra che la zona denuclearizzata proposta dalla commissione possa evitare quelle manchevolezze associate alle zone denuclearizzate relativamente alle armi da campo, proposte in precedenza. Che cosa si otterrebbe portando a termine l'attuazione di zone denuclearizzate relativamente alle armi da campo?

Tanto per cominciare, si rafforzerebbero le barriere contro l'avvio involontario, o accidentale, di una guerra nucleare. Sembra chiaro che, in tempi normali, una guerra nucleare non ha molte probabilità di scoppiare in base a un freddo calcolo di possibili vantaggi: le forze nucleari di entrambe le parti sono troppo grandi perché gli uni o gli altri possano pensare di ottenere vantaggi significativi. È più probabile che una guerra nucleare abbia origine da decisioni disperate, prese nel corso di una crisi acuta, una crisi che ha già visto impegnate, probabilmente, le forze convenzionali delle due parti. In circostanze del genere, l'una o l'altra delle due parti potrebbe dare il via a un attacco nucleare perché non vede nessun'altra possibilità per evitare un sconfitta catastrofica oppure perché ha tratto la conclusione che il proprio avversario si sta impegnando in una escalation del conflitto fino a giungere al livello nucleare.

La rimozione delle armi nucleari da una zona molto ampia dell'Europa centrale, situata ai due lati del confine Est--Ovest, ridurrebbe l'eventualità che la situazione faccia precipitare decisioni disperate di questo genere. Essa toglierebbe tali armi dal caos della guerra, per lo meno all'inizio, e darebbe ai leader di entrambe le parti un bene molto prezioso: il tempo. Permetterebbe un'analisi più attenta e ponderata della situazione militare tattica e la possibilità di continuare senza interruzione nell'uso esclusivo delle forze convenzionali. Ridurrebbe, negli uni e negli altri, la preoccupazione che la parte avversa ceda alle pressioni tendenti a farla ricorrere precipitosamente alle armi e attenuerebbe, di riflesso, la spinta a sferrare, per prima, un attacco. Permetterebbe, infine, una ricerca più accurata di alternative diplomatiche; in questo maggior tempo disponibile, prima di prendere la decisione di ricorrere alle armi nucleari, entrambe le parti potrebbero forse arrivare a rendersi conto dei pericoli eccezionali di cui è costellata la strada che intendono imboccare, e potrebbero prendere provvedimenti per risolvere il conflitto prima che esso sfugga loro di mano.

Se si attuassero le proposte della commissione per una zona senza armi nucleari da campo di battaglia, ci sarebbero anche altri vantaggi. I cambiamenti imposti da un accordo di questo genere potrebbero condurre entrambe le parti a riconsiderare in maniera autonoma gli attuali inventari delle armi nucleari, presenti sul continente. Molti osservatori hanno sostenuto che, sul piano militare, sono possibili notevoli riduzioni. Ciò che le ha impedite è stata la preoccupazione per le possibili implicazioni politiche di un ritiro unilaterale delle armi. Gli spostamenti, resi necessari dall'istituzione della zona denuclearizzata relativamente alle armi da campo e il fatto che la parte avversa stia procedendo ad analoghi cambiamenti potrebbero offrire l'occasione per ridurre notevolmente le forze nucleari senza che si abbiano conseguenze politiche negative.

Il successo del negoziato per una zona denuclearizzata relativamente alle armi da campo insieme con le misure proposte per un limite alle armi chimiche e convenzionali, schiuderebbe possibilità più ambiziose per controllare e per ridurre gli arsenali militari in Europa e in qualsiasi altra parte del mondo; esso avrebbe un effetto positivo sulle relazioni politiche fra nazioni dell'Occidente e dell'Est europeo, in quanto costituirebbe un primo passo sulla strada dei rapporti, meno tesi e più cooperativi, fra Stati Uniti e Unione Sovietica, e avvantaggerebbe anche i rapporti all'interno dell'alleanza occidentale, eliminando quella che è diventata ormai una rovinosa causa di attrito fra Stati Uniti ed Europa occidentale: la controversia sul modo di impostare il problema del controllo degli armamenti nucleari sul nostro continente.

Verso la fine dello scorso anno, il governo svedese ha sollecitato formalmente l'opinione di tutti i governi europei e nordamericani sulla proposta di una zona denuclearizzata relativamente alle armi da campo. Nel mese di gennaio, l'Unione Sovietica - come risposta - ha affermato di sottoscrivere l'idea e di avere l'intenzione di porre la proposta sul tavolo delle trattative ai colloqui di Vienna. Ha anche affermato però che quella proposta non andava abbastanza lontano, aggiungendo che la zona avrebbe dovuto avere una larghezza di 500-600 chilometri, il che includerebbe, praticamente, tutta la Germania Orientale e tutta la Germania Occidentale.

Fino a questo momento, l'Occidente ha accolto con freddezza l'idea di una zona denuclearizzata. Se, però, a Vienna i sovietici avanzassero davvero una proposta del genere, sarebbe un'occasione d'oro per legare esplicitamente l'attuazione di una zona denuclearizzata alla conclusione di un primo accordo sulla limitazione delle armi convenzionali e per negoziare, a proposito di tale zona, un'estensione che sia accettabile dai due blocchi e che non dia adito a problemi politici analoghi a quelli presentati dal piano Rapacki.

L'attuazione di una zona denuclearizzata relativamente alle armi da campo non eliminerebbe, in Europa, il rischio di una guerra nucleare. Un certo rischio rimarrà sempre fintantoché le armi nucleari continueranno a far parte degli arsenali delle maggiori potenze e il suo grado aumenterà o diminuirà in consonanza con le relazioni politiche generali e con i mutamenti specifici negli spiegamenti militari di entrambe le parti. È improbabile, invece, che il rischio scompaia del tutto: è una realtà con cui ognuno dovrà continuare a vivere. La proposta di una zona denuclearizzata relativamente alle armi da campo è, comunque, una misura pragmatica e politicamente possibile per controllare il pericolo di una guerra nucleare in Europa. Essa merita seria considerazione da parte di tutti i cittadini e di tutti i governi.

Dispositivi micromeccanici in silicio

Mediante incisione chimica di un wafer di silicio monocristallino, si possono realizzare piccolissimi dispositivi meccanici, di cui è possibile la produzione in grande serie, come per i circuiti microelettronici

di James B. Angell, Stephen C. Terry e Phillip W. Barth

Negli ultimi trent'anni il silicio è divenuto familiare come materiale con cui si realizzano componenti e sistemi elettronici. Sotto forma di chip microelettronici si ritrova in una vastissima gamma di prodotti che vanno dalla lavastoviglie ai sistemi di controllo dello *Space Shuttle.* Lo sfruttamento delle caratteristiche elettriche del silicio è stato accompagnato da una ricerca meno pubblicizzata sulle altre sue proprietà e sul loro impiego potenziale, che ha condotto allo sviluppo di una tecnologia detta microlavorazione che consente di realizzare dispositivi meccanici in silicio piccoli quasi quanto quelli microelettronici.

La microlavorazione inizia con le medesime tecniche di fabbricazione che hanno reso economici i chip di silicio per i circuiti integrati. Si tratta di processi che consentono di ottenere numerosi chip allo stesso tempo con conseguente ripartizione del costo di produzione e abbassamento del costo unitario del prodotto. La microlavorazione aggiunge a tali tecniche l'incisione chimica per foggiare forme tridimensionali come cavità, fori, piramidi, canali, semisfere, elementi a sbalzo, diaframmi, aghi e pareti. Combinando questi elementi strutturali è possibile realizzare numerosi dispositivi meccanici.

Si sono prodotti vari micromeccanismi in silicio: valvole, molle, specchi, ugelli, raccordi, piastre per circuiti, testine per stampanti, dissipatori di calore e trasduttori di grandezze quali forza, pressione, accelerazione e concentrazione chimica. Perfino un dispositivo complesso come un gascromatografo, uno strumento che è in grado di eseguire l'analisi qualitativa e quantitativa di una miscela di composizione ignota, può essere costruito su un disco di silicio del diametro di pochi centimetri.

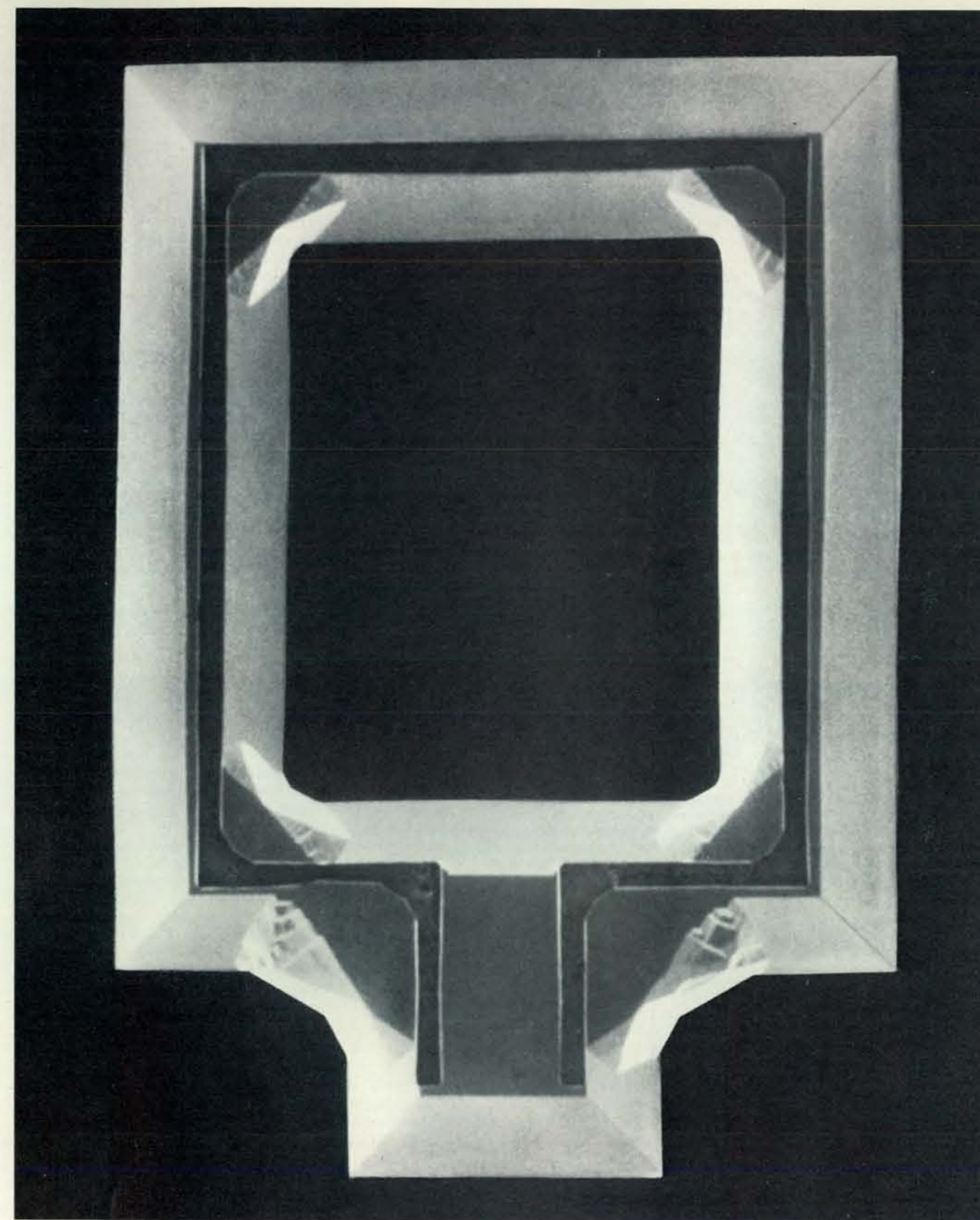
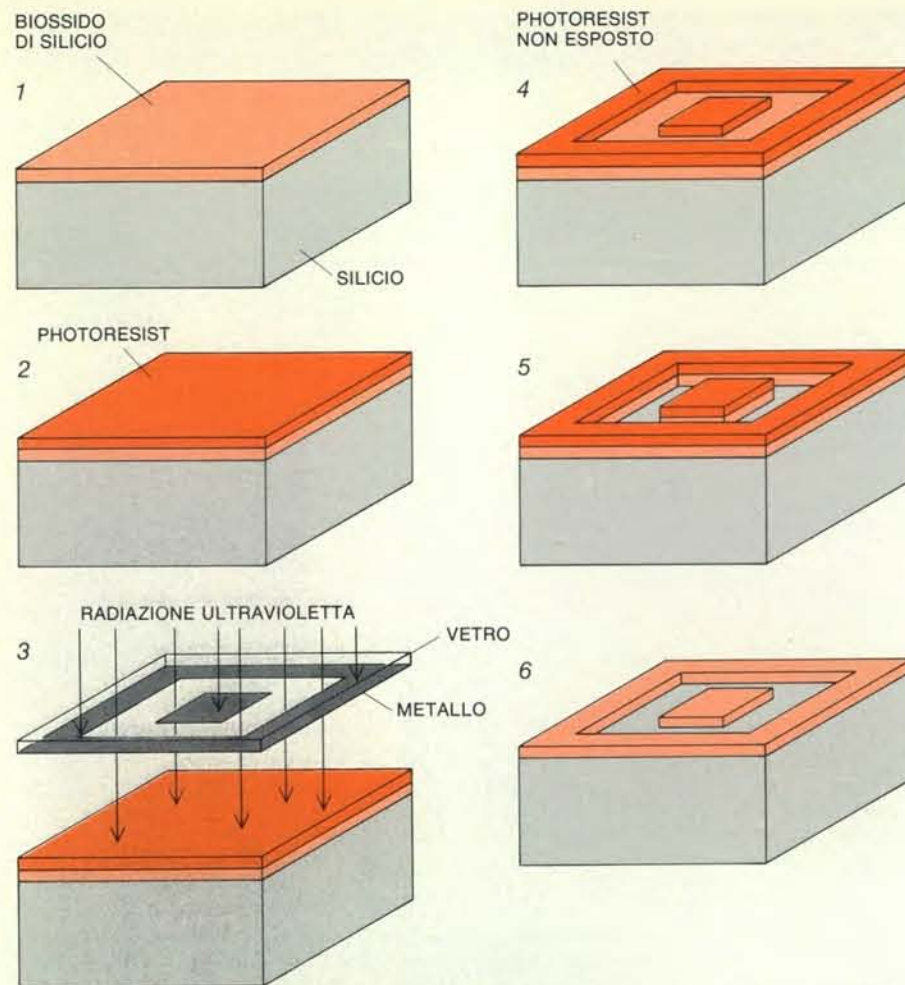
Gran parte dell'interesse per la microlavorazione deriva dalla necessità di trasduttori più economici e di maggiore versatilità. Fino a non molto tempo addietro i componenti elettronici dei sistemi automatici di controllo e degli strumenti di misura erano più costosi dei trasduttori, ma con l'introduzione del microelaboratore il costo dell'elettronica ha iniziato a diminuire rapidamente. Trasduttori in silicio fabbricati in grande serie potrebbero sostituire i trasduttori costruiti a mano, più o meno come i circuiti in silicio fabbricati in grande serie hanno sostituito le valvole termoioniche.

L'ultimo sviluppo nel settore è il trasduttore integrato, un chip in silicio che comprende sia il trasduttore vero e proprio sia il circuito elettronico connesso, necessario per trattare il segnale ottenuto. I trasduttori integrati sono meno costosi dell'insieme costituito dai trasduttori e dai necessari componenti elettronici fabbricati separatamente e inoltre, grazie alla minore sensibilità nei confronti del rumore e delle dispersioni di corrente, hanno prestazioni superiori ai sistemi discreti e consentono una gamma più ampia di applicazioni.

Il silicio è un semiconduttore ed è uno degli elementi che nel sistema periodico si trovano fra i metalli e i non metalli. Si differenzia però dagli altri semiconduttori per il fatto di essere facilmente ossidabile: esposto al vapore forma in superficie uno strato di biossido (SiO_2), inerte dal punto di vista chimico e isolante da quello elettrico, che, per quanto riguarda la struttura, risulta sostanzialmente un vetro. Gli strati di ossido sono sfruttati per proteggere aree di silicio durante la fabbricazione di dispositivi microelettronici: in ciò risiede il motivo del largo impiego di questa sostanza anziché di altri semiconduttori quali il germanio nella fabbricazione dei circuiti integrati. Anche nella microlavorazione si mettono attivamente a profitto le possibilità offerte dagli strati superficiali di biossido di silicio.

Nell'industria elettronica si usano monocristalli di silicio la cui produzione è già di per se stessa una raffinata tecnologia. Grossi monocristalli o «*boule*» dell'elemento, aventi diametro di 10 centimetri e lunghezza di un metro, vengono tagliati in lamine *(wafer)* a forma di disco, di spessore compreso fra 0,2 e 0,5 millimetri. I wafer sono poi lucidati a specchio per eliminare i difetti introdotti dal taglio. La struttura cristallina omogenea impartisce al materiale le proprietà elettriche richieste per i circuiti microelettronici. Si è visto, però, che in questa forma l'elemento possiede anche particolari proprietà meccaniche.

Il silicio, che nella tavola periodica degli elementi si trova direttamente al disotto del carbonio, forma cristalli dello stesso tipo del diamante, anche se i legami interatomici sono alquanto più deboli. I monocristalli di silicio sono fragili e possono venire sfaldati come i diamanti, ma sono più duri della maggior parte dei metalli. Inoltre il silicio monocristallino resiste in modo sorprendente alle sollecitazioni meccaniche: i limiti di deformazione elastica alla trazione e alla compressione sono superiori a quelli dell'acciaio. Di contro, raggiunto il limite, il silicio si frattura mentre l'acciaio subisce una deformazione permanente. Infine il silicio monocristallino, sottoposto a ripetuti cicli di trazione e compressione, conserva la propria resistenza, mentre i metalli policristallini sono soggetti a rotture per fatica, tendono cioè a indebolirsi e a fratturarsi perché le sollecitazioni si concentrano lungo le superfici intercristalline.

La microlavorazione e la fabbricazione microelettronica iniziano con la fotolitografia, la tecnica fotografica usata per riportare la figura di una matrice sulla superficie del wafer di silicio. La prima fase del processo fotolitografico consiste nel far formare un sottile strato di ossido sulla superficie del wafer mediante riscaldamento fra 800 e 1200 gradi centigradi in atmosfera di vapore (può essere usato anche ossigeno secco, ma il vapore agisce molto più in fretta). Nella fase successiva si deposita sulla superficie di ossido un sottile strato di un polimero organico sensibile alla radiazione ultravioletta, chiamato photoresist. A contatto della super-

L'accelerometro al silicio sviluppato da Lynn M. Roylance della Stanford University è uno strumento miniaturizzato che sfrutta sia le proprietà elettriche sia quelle meccaniche del silicio. Esso è ottenuto mediante incisione chimica selettiva di un «wafer» (cioè una lamina), tagliato da un grosso monocristallo di silicio. L'illustrazione è una microfotografia, ottenuta al microscopio elettronico a scansione, dell'accelerometro, costituito essenzialmente da una massa di silicio avente uno spessore uguale a quello del wafer, sospesa all'estremità di un sottile elemento, sempre in silicio. Nella parte inferiore di quest'ultimo è ricavato un resistore elettrico, la cui resistenza varia quando l'elemento si flette, consentendo così di misurare l'accelerazione. La massa viene liberata dal materiale circostante mediante attacco chimico. L'elemento viene realizzato eliminando con lo stesso procedimento tutto il materiale a esclusione di un sottile strato di silicio in quella zona del wafer. Il dispositivo impaccato ha una superficie di due per tre millimetri con uno spessore di 0,6 millimetri ed è quindi così piccolo da poter essere suturato al cuore per misurarne l'accelerazione. Da un wafer di silicio possono essere ottenuti contemporaneamente centinaia di accelerometri.

Sulla superficie del wafer una figura di ossido protegge aree opportunamente prescelte durante l'attacco chimico a umido con cui si realizza un dispositivo micromeccanico. Si inizia con la formazione di uno strato di biossido di silicio (SiO_2) sulla superficie del wafer; il silicio reagisce con il vapore a temperature comprese fra 800 e 1200 gradi centigradi ossidandosi. Il wafer ossidato è quindi rivestito con un polimero sensibile alla radiazione ultravioletta denominato photoresist. Successivamente il wafer viene posto sotto una maschera madre (una figura metallica su una piastra di vetro) e sottoposto a radiazione. Il photoresist nelle aree esposte si indebolisce ed è quindi asportato in una soluzione di sviluppo. Il wafer viene poi attaccato con un acido che intacca l'ossido non ricoperto, ma non il photoresist residuo, né il silicio. Si ottengono in questo modo una serie di aperture nell'ossido che riproducono la figura metallica sulla piastra di vetro. Nelle fasi successive del processo la figura di ossido funge da maschera per incisione del silicio sottostante.

ficie ricoperta dal photoresist si pone una fotomaschera, costituita in genere da una lastra di vetro su cui è riportata una figura metallica, poi si espone il wafer a radiazione ultravioletta. Alle lunghezze d'onda dell'ultravioletto il metallo della maschera è opaco e il vetro trasparente: la radiazione provoca una reazione chimica nell'area esposta del photoresist; se questo è del tipo detto positivo, la reazione chimica indebolisce il polimero, mentre se è del tipo negativo lo rinforza. Dopo tale operazione il wafer viene lavato in una soluzione di «sviluppo» che asporta l'area esposta oppure quella non esposta del photoresist, lasciando una figura di photoresist e di ossido nudo sulla superficie del wafer. La figura di photoresist costituisce un'immagine positiva o negativa della figura riportata sulla maschera.

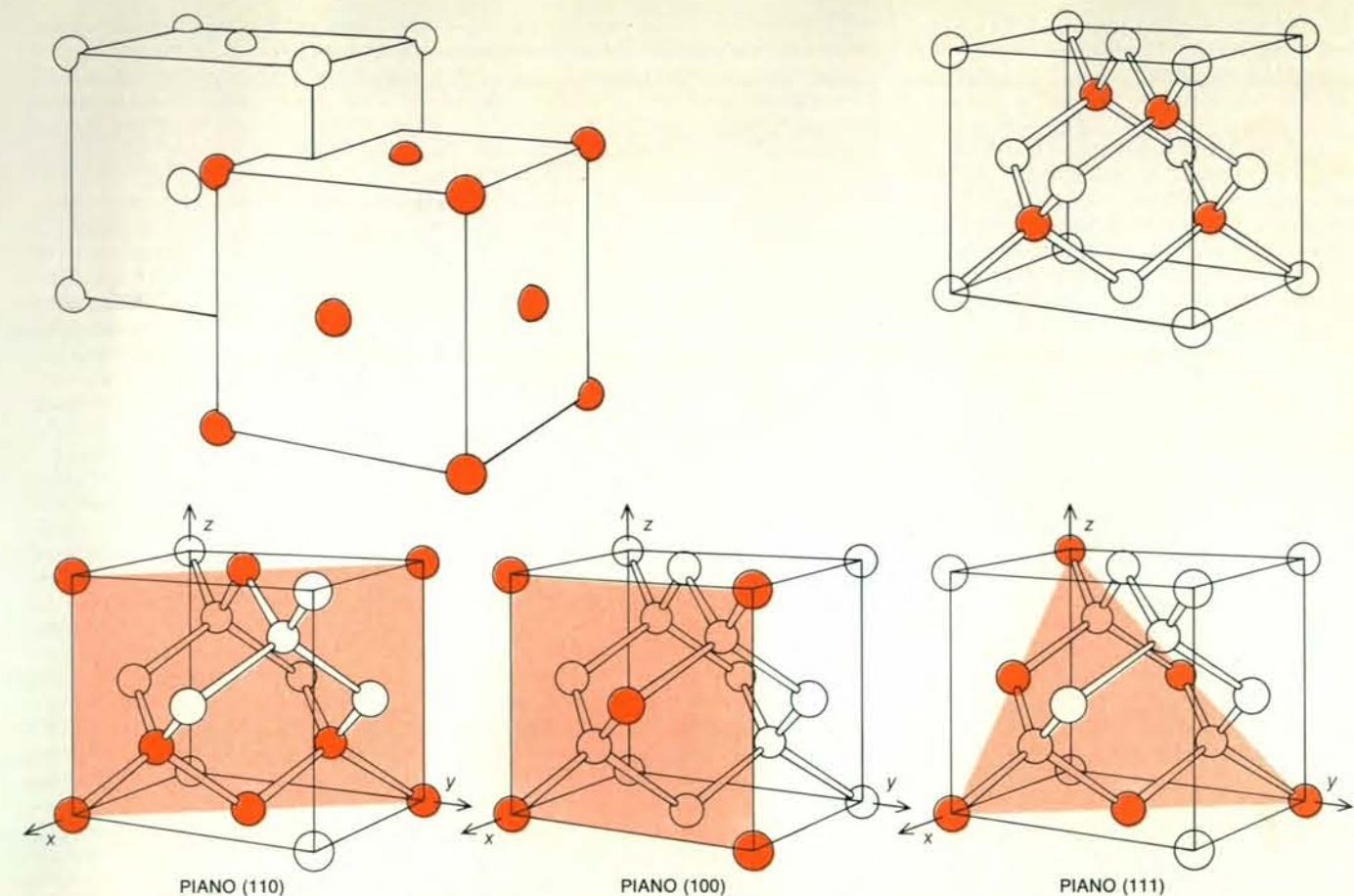
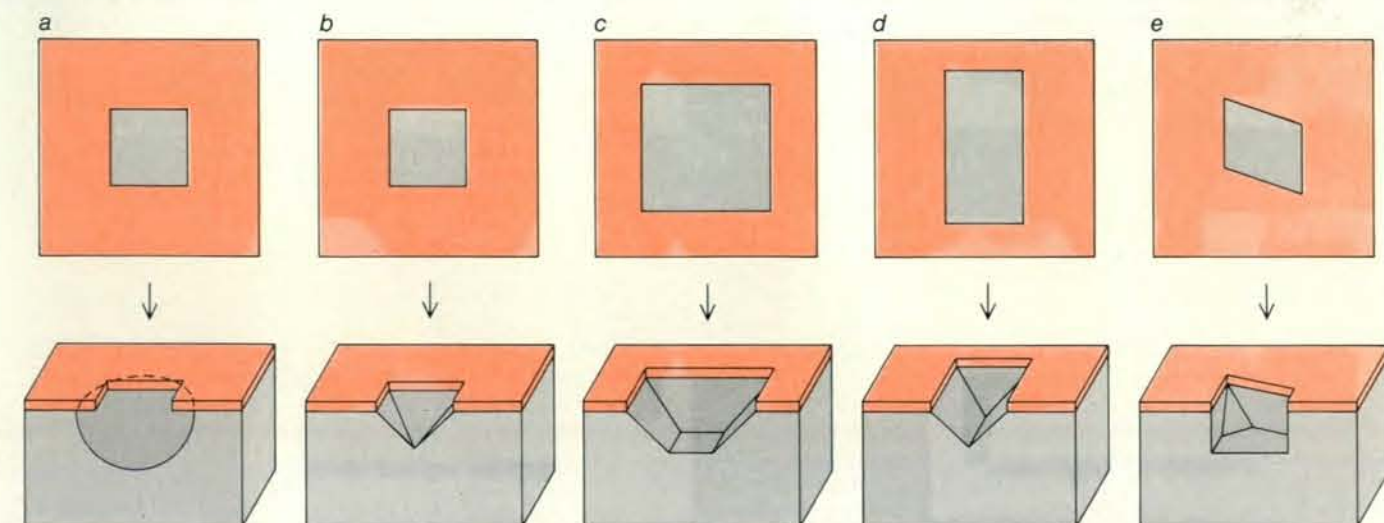
Si pone poi il wafer in una soluzione di acido fluoridrico che attacca l'ossido, ma non il silicio sottostante, né il photoresist, che protegge le aree ossidate che ricopre. Dopo la rimozione dell'ossido esposto, si può togliere il photoresist restante con acido solforico caldo che attacca il photoresist stesso, ma non l'ossido né il silicio. Il risultato che si ottiene al termine è una figura di ossido sulla superficie del wafer che riproduce la figura del photoresist ed è quindi una copia positiva o negativa della figura della maschera. La figura di ossido funge a sua volta da maschera nelle fasi successive del procedimento.

A questo punto la microlavorazione procede in modo diverso dalla fabbricazione microelettronica. Nel trattamento di un dispositivo microelettronico la figura di ossido funge da maschera durante il «drogaggio» del wafer con impurezze (per esempio boro o fosforo). Dopo il riscaldamento del wafer, gli atomi di impurezze depositati sulla superficie diffondono per un piccolo spessore nel silicio non ricoperto da ossido, creando regioni conduttive o resistive di piccolo spessore che compongono il dispositivo microelettronico. Nella microlavorazione l'ossido è usato quale maschera durante l'attacco chimico. Gli agenti attaccano il silicio esposto dalle aperture nello strato di ossido e scavano profonde cavità tridimensionali nel wafer.

Nella microlavorazione del silicio si impiegano due tipi di agenti: gli agenti isotropi incidono il cristallo di silicio alla stessa velocità in tutte le direzioni creando forme lievemente arrotondate; quelli anisotropi, noti anche come agenti cristallografici o agenti dipendenti dall'orientamento, incidono a velocità differenti in direzioni diverse il reticolo cristallino, dando luogo così a forme bene definite con angoli e bordi netti. Gli agenti isotropi più utili sono miscele di acidi fluoridrico, nitrico e acetico (agenti FNA). Gli agenti anisotropi sono costituiti da soluzioni alcaline che agiscono a caldo: idrossido di potassio (KOH), idrossido di sodio (NaOH) e una miscela, composta da etilendiammina, pirocatechina (*o*-diossibenzene) e acqua, nota come EDP.

La microlavorazione usa di preferenza gli agenti anisotropi rispetto a quelli isotropi. Una tipica operazione di microlavorazione, l'incisione nel wafer di una schiera di fori passanti, ne chiarirà il motivo. Partendo dalle aperture nell'ossido, un agente isotropo agisce sia in basso sia di lato e attacca anche le zone al disotto della maschera di ossido allargando la cavità mentre la approfondisce. La cavità così ottenuta è larga almeno il doppio della sua profondità e quindi la distanza fra le aperture nello strato di ossido deve essere almeno pari allo spessore del wafer, per poter ottenere fori passanti separati. Di contro un agente anisotropo, impiegato in modo appropriato, non agisce sotto la maschera di ossido e crea cavità con pareti laterali ben definite, cioè che non si allargano mentre si approfondiscono. Perciò un agente anisotropo può creare schiere di fori molto ravvicinati fra loro e i bordi delle aperture nella maschera di ossido possono essere ravvicinati tanto quanto consentono i limiti del processo fotolitografico.

La forma del foro prodotto da un agente anisotropo dipende dall'orientamento dei piani atomici nel wafer. Come è possibile identificare differenti file unidimensionali di steli in un campo bidimensionale di mais osservandolo sotto angoli diversi, così in un cristallo tridimensionale diversi piani bidimensionali di atomi sono disposti ad angoli differenti fra loro. I piani di atomi nel cristallo vengono identificati mediante un sistema di coordinate assegnate al cristallo.

Ogni atomo di silicio è legato a quattro atomi circostanti: la struttura di tale legame è tetraedrica, ossia ogni atomo si trova al centro di un tetraedro definito dai quattro atomi cui è legato. Non è conveniente però assumere la struttura del legame quale base di un sistema di coordi-

Il silicio possiede la medesima struttura cristallina del diamante, nota come struttura cubica a facce centrate. A ogni vertice e al centro di ogni faccia del cubo unitario si trova un atomo, mentre i cubi sono concatenati in modo tale che diversi atomi appartenenti a cubi circostanti giacciono all'interno di ogni cubo unitario. Gli assi del cubo unitario formano un sistema di coordinate rette che rende possibile definire direzioni e piani all'interno del cristallo. Una direzione cristallina viene individuata da tre coordinate dette indici di Miller, che sono multipli interi della lunghezza di uno spigolo del cubo unitario. Lo stesso gruppo di indici individua i piani perpendicolari alla direzione. L'orientamento cristallino ha un ruolo importante nella fabbricazione dei dispositivi micromeccanici perché alcuni agenti hanno velocità di incisione diverse: la maggior parte degli agenti anisotropi progredisce velocemente nella direzione cristallina normale al piano (110) e meno velocemente nella direzione perpendicolare al piano (100). L'incisione in direzione normale al piano (111) procede a velocità bassissima o addirittura nulla.

Gli agenti anisotropi creano fori sfaccettati composti dai piani cristallini che vengono attaccati più lentamente, mentre gli agenti isotropi creano sempre fori leggermente arrotondati *(a)*. La forma di un foro ottenuto con un agente anisotropo è determinata dall'orientamento cristallino della superficie del wafer, dalla forma e dall'orientamento delle aperture della maschera sulla superficie stessa e dalla dipendenza dall'orientamento dell'agente di attacco. Un'apertura quadrata orientata secondo le direzioni <110> di un wafer <100> da luogo a cavità piramidali con pareti laterali {111} *(b)*. Quanto più larga è l'apertura tanto più profondo è il punto di intersezione dei piani {111}, ed è in questo modo possibile creare cavità a fondo piano arrestando l'incisione prima di raggiungere tale profondità *(c)*. Un'apertura rettangolare sullo stesso wafer produce un solco a forma di *V (d)*. Si possono creare fori con pareti laterali parallele incidendo un wafer che ha un differente orientamento superficiale. In un wafer <110> due serie di piani {111} sono perpendicolari alla superficie, ma non fra di loro. Se l'apertura nell'ossido di un wafer <110> è orientata in modo opportuno, l'incisione crea un foro con pareti laterali verticali. Le pareti laterali che si intersecano ad angolo acuto sono attraversate anche da altri piani. Gli esempi nell'illustrazione sono solo alcune delle molte forme ottenibili per attacco chimico.

nate poiché in essa i legami atomici non sono disposti ad angolo retto fra loro. È meglio invece immaginare che il cristallo di silicio sia costituito da strati sovrapposti di cubi tutti uguali. Ogni cubo ha un atomo posto in corrispondenza di ciascun vertice e al centro di ogni faccia ed è concatenato con quattro cubi adiacenti. Tale struttura cubica a facce centrate può servire come base per un sistema di coordinate. Ogni atomo può essere preso come origine del sistema di coordinate i cui assi giacciono lungo gli spigoli del cubo al quale appartiene l'atomo. L'unità di misura è la lunghezza dello spigolo di un cubo.

Un vettore, cioè una direzione nel cristallo, è individuato da tre coordinate, chiamate indici di Miller. Per esempio un vettore [110] punta diagonalmente attraverso la faccia di un cubo unitario. Le coordinate di un dato vettore definiscono anche l'intera serie di piani atomici perpendicolari a quella direzione. Così la notazione (100) descrive tutti i piani normali all'asse *x*. Poiché la struttura del cristallo è simmetrica, le direzioni *x*, *y* e *z* sono intercambiabili fra loro ed è possibile introdurre una notazione generalizzata per descrivere tutte le direzioni e le serie di piani equivalenti. La notazione <110> individua le diagonali di qualsiasi faccia di un cubo unitario e la notazione {110} individua la serie di piani perpendicolari a tutti i vettori <110>. Le parentesi di tipo diverso distinguono i piani dalle direzioni e piani e direzioni generici da piani e direzioni particolari.

I meccanismi che sono alla base della relazione che lega la velocità di attacco di un agente anisotropo con le direzioni cristalline non sono ben noti. Le differenze fra le velocità di attacco dipendono dalla temperatura e dalla composizione dell'agente, a loro volta apparentemente funzione delle densità dei legami atomici nel piano esposto di silicio e del raggio degli ioni ossidrile nella soluzione di attacco.

L'esperienza ha dimostrato comunque che i piani {111} esposti vengono attaccati lentamente (quando ciò avviene) da parte degli agenti anisotropi. Al contrario i piani {110} in genere vengono attaccati rapidamente e quelli {100} a una velocità intermedia. Fori semplici ottenuti per attacco possono presentare solo sfaccettature {110} e {100}. Altri piani di ordine superiore, come i piani {221}, vengono attaccati a una velocità intermedia e possono apparire come sfaccettature di forme complesse.

Il tipo di foro formato in un wafer da un agente anisotropo è determinato dall'orientamento cristallino della superficie del wafer come pure dalla dipendenza dall'orientamento dell'agente. I due tipi di wafer in silicio che risultano più utili per l'incisione da parte di agenti anisotropi sono i wafer <100> (nei quali i piani {100} sono paralleli alla superficie e una direzione <100> è perpendicolare a quest'ultima) e i wafer <110> (nei quali i piani {110} sono paralleli alla superficie e una direzione <110> è perpendicolare a quest'ultima).

Anche l'orientamento, la dimensione e la forma delle aperture di ossido sulla superficie del wafer hanno un ruolo nel determinare il tipo di foro che si ottiene, come mostrerà un esempio. Se su un wafer <100> si ha un'apertura quadrata nello strato di ossido, il trattamento con un agente anisotropo può creare una cavità piramidale con vertice verso il basso. La superficie di un wafer <100> è intersecata da quattro famiglie di piani {111} lungo le due direzioni normali <110> che giacciono nel piano della superficie. Durante l'attacco anisotropo il piano esposto della superficie {100} viene attaccato verso il basso con velocità costante, con il risultato di produrre una cavità con fondo piano all'inizio del processo. Quattro piani {111} inclinati verso l'interno (ad angoli di 55 gradi con la superficie) formano le pareti della cavità partendo dai bordi del foro. Continuando l'attacco risultano esposte maggiori superfici dei piani {111} e l'area del fondo piano si restringe. Alla fine i piani {111} si incontrano in un certo punto e il fondo piano scompare, mentre l'incisione termina per mancanza di piani esposti facilmente intaccabili.

L'apertura quadrata nell'ossido deve essere allineata con precisione (la tolleranza è di uno o due gradi) con le direzioni <110> sulla superficie del wafer al fine di ottenere cavità piramidali corrispondenti con esattezza alla maschera di ossido, senza asportazione di materiale sotto la medesima. La maggior parte dei wafer di silicio ha un bordo parallelo alla direzione <110> nel wafer stesso. Il tracciato quadrato della maschera è allineato con tale bordo durante la fase fotolitografica che precede l'incisione.

La dimensione della figura di ossido determina non solo l'area della cavità sulla superficie del wafer, ma anche la sua profondità. Quanto maggiore è l'apertura quadrata, tanto più in profondità è il punto in cui le pareti laterali {111} della cavità si intersecano. Se l'apertura nell'ossido è sufficientemente grande, i piani {111} non si intersecano all'interno del wafer e quindi la cavità diventa passante con la creazione di un piccolo foro quadrato sulla superficie di fondo.

L'influenza della forma dell'apertura nell'ossido sulla geometria della cavità incisa è semplice. Se l'apertura nell'ossido su un wafer <100> è rettangolare anziché quadrata, l'incisione forma un solco anziché una piramide rovesciata. Le pareti e le estremità del solco sono piani {111} e, se l'incisione procede fino in fondo, il solco presenta una sezione trasversale a forma di *V*.

L'ugello in silicio per stampante a getto d'inchiostro ad alta velocità è un foro piramidale il cui fondo è costituito da una sottile membrana nella quale è stato inciso chimicamente un orifizio di diametro ben preciso. Tale diametro è determinato dalle prime tre fasi del processo. Il wafer viene ossidato e sulla superficie superiore si depositano dei puntini di photoresist. Si asporta quindi l'ossido non protetto dal photoresist lasciando una serie di punti di ossido. Facendo diffondere del boro nel silicio si crea dappertutto, salvo sotto i punti di ossido, uno strato di silicio drogato con boro. Quindi si crea un'apertura quadrata nell'ossido della superficie inferiore del wafer, che viene inciso con un agente anisotropo, il quale forma una cavità piramidale con pareti laterali {111}; l'incisione si arresta quando perviene allo strato drogato, ma continua attraverso i punti non drogati nello strato di modo che, quando tutto l'ossido è eliminato, si realizzano gli orifizi passanti degli ugelli. Infine il wafer viene ossidato di nuovo per fornire a tutte le superfici uno strato protettivo. Sullo stesso wafer si ottengono molte schiere di ugelli.

Cavità piramidali, solchi a forma di *V* e altre cavità e fori sfaccettati sono un tipo di elemento strutturale di cui sono costituiti i dispositivi micromeccanici. Una seconda unità strutturale è una sottile membrana di silicio. È possibile ottenere una struttura del genere incidendo un wafer per un periodo appena inferiore a quello occorrente per ottenere un foro passante. Con tale metodo è difficile però realizzare diaframmi di spessore uniforme, poiché lo spessore del diaframma è determinato da quello del wafer, che varia tipicamente di 10 micrometri o più in uno stesso wafer e da un wafer all'altro.

Si possono ottenere membrane di spessore e forma definiti con maggiore precisione mediante una tecnica che sfrutta un'altra proprietà degli agenti anisotropi. La velocità a cui le soluzioni incidono un wafer dipende non solo dall'orientamento cristallino, ma anche dal maggiore o minore drogaggio del silicio con atomi di impurezze quali boro, fosforo, arsenico o antimonio.

Se l'azione di un agente dipende dal drogaggio (il che non ha nulla a che vedere con la sua isotropia o meno), la sua velocità di incisione è funzione del tipo degli atomi droganti e della loro concentrazione. L'agente isotropo FNA dipende dal drogaggio per certi rapporti di miscela: attacca il silicio fortemente drogato molto più rapidamente del silicio poco drogato. Al contrario gli agenti anisotropi idrossido di potassio ed EDP attaccano molto più lentamente il silicio fortemente drogato con boro che non quello leggermente drogato con boro.

La prima fase per ottenere una membrana con l'uso di un agente anisotropo è la creazione di uno strato di arresto dell'incisione costituito da silicio fortemente drogato con boro sulla superficie del wafer. Sulla superficie anteriore del wafer viene depositato uno strato di una sostanza ricca di boro, riscaldando poi il tutto in un forno a una temperatura compresa fra 1000 e 1200 gradi centigradi. A queste temperature gli atomi di boro diffondono nel silicio. La profondità dello strato drogato è controllata dalla temperatura e dalla durata della fase di diffusione. Se poi il wafer viene attaccato dalla parte posteriore con un agente anisotropo, l'agente rimane attivo fino a quando raggiunge lo strato drogato creando una membrana spessa quanto lo strato medesimo sulla superficie del wafer.

Se il progettista vuole ottenere una membrana a qualche distanza dalla superficie del wafer, il processo prevede un analogo drogaggio della superficie e la successiva creazione di uno strato di silicio, al disopra di quello drogato, mediante il riscaldamento del wafer in atmosfera di silano gassoso (SiH_4). Alle alte temperature il gas si decompone depositando silicio sulla superficie del wafer. Il nuovo strato, denominato strato epitassiale, assume la struttura e l'orientamento cristallino del silicio su cui è depositato. Lo spessore tipico di uno strato epitassiale è compreso fra 5 e 20 micrometri.

Se il wafer viene poi attaccato da un agente anisotropo su entrambe le facce, si crea una membrana a una profondità pari allo spessore dello strato epitassiale. Lo spessore della membrana stessa è a sua volta determinato dallo spessore dello strato drogato con boro.

Un terzo esempio di struttura micromeccanica è rappresentato dal dispositivo a sbalzo, un sottile elemento di biossido di silicio, sostenuto a una sola estremità. In molte applicazioni un elemento a sbalzo di questo tipo è sospeso sopra una cavità poco profonda realizzata nel silicio. I procedimenti per ottenere tale struttura sono stati sviluppati da un gruppo che lavorava con Kurt E. Petersen al San Jose Research Laboratory della International Business Machines Corporation, sfruttando due proprietà degli agenti anisotropi: la dipendenza dal drogaggio e la tendenza a scavare forme convesse scalzando il materiale sotto una maschera di ossido.

Le semplici aperture geometriche nell'ossido, finora considerate, sono forme concave, cioè la linea che separa l'ossido e l'apertura è sempre concava rispetto

PRIMA DELL'INCISIONE

10 MINUTI

40 MINUTI

80 MINUTI

Si può formare un elemento a sbalzo scavando il materiale sotto una maschera di ossido. Un agente anisotropo scava al di sotto delle forme convesse di ossido, mentre lascia intatte quelle concave opportunamente orientate. Nell'illustrazione l'apertura rettangolare nell'ossido è una forma concava; le sporgenze di ossido che si protendono nell'apertura sono forme convesse. Se l'avanzamento dell'agente verso il basso viene fermato da uno strato di arresto, un attacco prolungato con agente anisotropo dà luogo a una linguetta di ossido sospesa su una cavità rettangolare con pareti laterali {111} e a un fondo piano costituito dallo strato di arresto dell'incisione. Elementi a sbalzo, del tipo di quelli che si ottengono quando la sporgenza di ossido è completamente scalzata, sono stati usati per realizzare accelerometri, schiere di interruttori controllati per via elettrostatica e schiere di specchi. Questo tipo di scavo rende difficile incidere un foro nel quale sporge un bordo di silicio ad angolo retto: il silicio sotto la sporgenza di ossido ad angolo retto in basso a sinistra è rimasto arrotondato durante l'incisione. Sotto la sporgenza circolare di ossido in alto a destra, durante l'incisione si forma un angolo nel silicio di circa 90 gradi, che viene però eroso se l'incisione continua. Si può interrompere l'incisione, in modo da ottenere fori in cui si protendano angoli retti al di sotto di forme circolari.

all'apertura stessa. Una forma concava di ossido non viene mai scalzata da un agente anisotropo se l'apertura è opportunamente orientata sulla superficie del wafer. Tuttavia se da una parte dell'apertura l'ossido forma un dente o una sporgenza, la sua periferia risulta convessa rispetto al silicio esposto. Un agente anisotropo attacca anche al di sotto di tali aperture convesse.

La prima fase per realizzare un elemento a sbalzo su una cavità poco profonda è la formazione di uno strato di arresto dell'incisione drogato con boro. Sopra lo strato drogato si deposita uno strato epitassiale di silicio, che viene ossidato; nell'ossido si pratica infine un'apertura. Il wafer è poi attaccato con un agente anisotropo. Se nell'apertura sporge un dente di ossido, l'agente comincia a intaccarlo anche al di sotto. Quando l'agente in corrispondenza dell'estremità del dente in contatto con l'ossido perviene a un piano {111} l'incisione ha termine. Poiché lo strato di arresto ferma l'avanzamento dell'incisione, il risultato finale è un elemento di ossido disposto a sbalzo su una depressione poco profonda nel substrato.

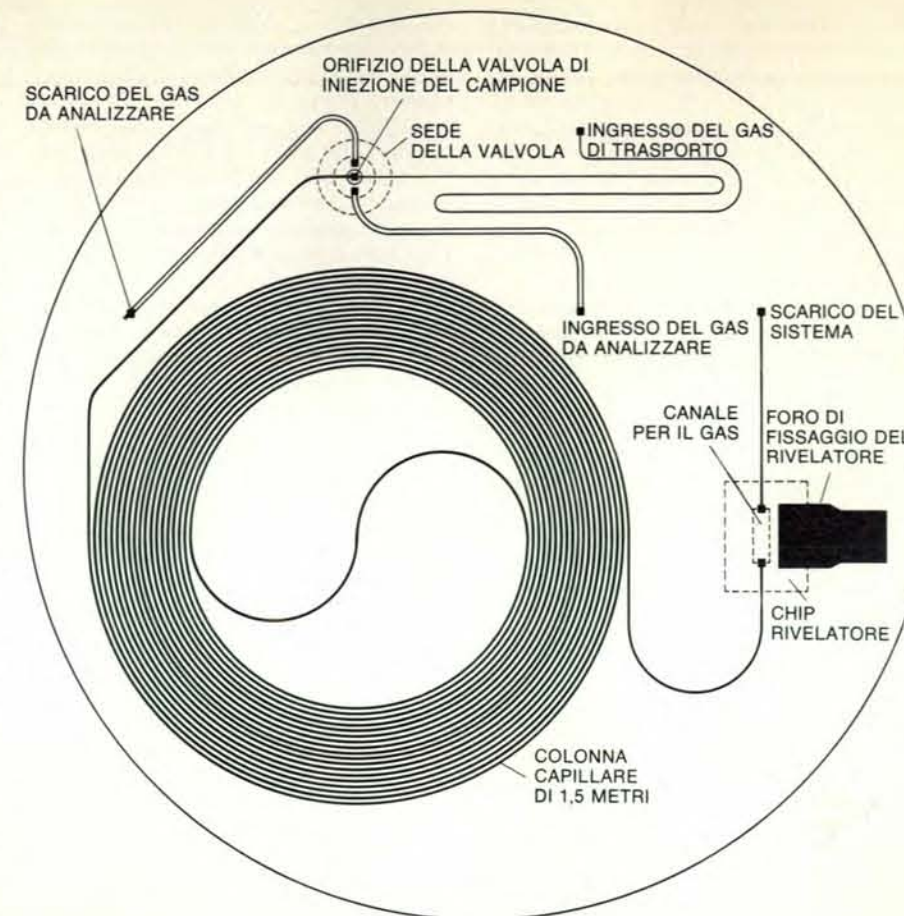
Presso l'Integrated Circuits Laboratory della Stanford University il lavoro sulla microlavorazione iniziò nel 1965 quando Kensall D. Wise (ora all'Università del Michigan) e uno di noi (Angell) svilupparono una fitta schiera di elettrodi per registrare i potenziali elettrici nel cervello. Il progetto condusse allo sviluppo di vari tipi di sonde per il cervello. Da allora i lavori sulla microlavorazione a Stanford hanno interessato pressosensori, accelerometri, gascromatografi, schiere di termometri miniaturizzati e strutture per l'isolamento dielettrico di circuiti microelettronici. Diversi altri laboratori universitari e industriali stanno sviluppando tecniche e dispositivi micromeccanici: quelli descritti nel seguito dell'articolo sono stati sviluppati in parte dagli autori e in parte da altri ricercatori.

Il gascromatografo in silicio è uno strumento analitico costruito su un wafer di cinque centimetri di diametro. La colonna capillare del cromatografo, una spirale lunga 1,5 metri, è costituita da un solco ottenuto per incisione del wafer sul quale è fissato poi un vetro. Il prototipo è stato sviluppato presso la Stanford University con un finanziamento del National Institute for Occupational Safety and Health degli Stati Uniti.

Schiere di ugelli realizzate in silicio hanno trovato applicazione nelle stampanti a getto d'inchiostro per calcolatori digitali. In un tipo di stampante l'inchiostro viene iniettato a pressione attraverso una schiera lineare di ugelli ottenendo diversi getti paralleli, che all'uscita si suddividono in goccioline. Le goccioline di ogni getto passano in successione attraverso un elettrodo che, secondo il proprio stato, impartisce a ciascuna di esse una carica, oppure le lascia neutre. Una goccia neutra passa attraverso un campo elettrostatico senza deviare dalla propria traiettoria e lascia un segno sulla carta, mentre una goccia carica viene deflessa dal campo e, anziché colpire la carta, viene intercettata da un canale di raccolta. Un carattere viene formato selezionando le gocce emesse da una schiera verticale di ugelli mentre l'intero meccanismo si muove in direzione orizzontale davanti alla carta.

Ernest Bassous, Larry Kuhn e i loro colleghi presso il Thomas J. Watson Research Center della IBM hanno condotto le ricerche iniziali sulle schiere di ugelli micromeccanici per iniettori di inchiostro, sviluppando molti sistemi differenti. Il metodo più semplice per ottenere un ugello è quello di incidere una cavità piramidale nel silicio. Se il punto di intersezione dei piani {111} che formano le pareti della cavità si trova al di sotto della superficie inferiore del wafer, si crea una piccola apertura quadrata. Con questa tecnica è difficile ottenere orifizi di dimensioni uniformi, perché le loro dimensioni dipendono anche dallo spessore del wafer in quel dato punto e, come notato più sopra, tale spessore è difficile da controllare.

Il problema può essere risolto, per esempio, con l'aggiunta di una membrana dello spessore di tre micrometri sul fondo di ogni cavità destinata a fungere da ugello. Una cavità piramidale viene quindi ricavata mediante incisione fino alla membrana, ma l'orifizio dell'ugello è definito da un procedimento di diffusione del boro e non dalla geometria della cavità. Con questo metodo Bassous e i suoi collaboratori hanno realizzato schiere di otto ugelli con orifizio circolare aventi tutti un diametro di 20 micrometri esatti e distanziati fra loro esattamente di 0,3 millimetri.

La prima fase per ottenere la schiera è di definire gli orifizi degli ugelli come puntini di photoresist sulla superficie superiore di un wafer ricoperto di ossido dalle due parti. Si elimina poi dalla superficie tutto l'ossido salvo quello protetto dal photoresist. La superficie viene quindi drogata con boro che è fatto penetrare nel silicio per uno spessore di tre micrometri, mediante riscaldamento dei wafer in un forno di diffusione. Solo il silicio che si trova sotto i puntini di photoresist rimane non drogato.

In un gascromatografo la separazione dei gas si basa sulla loro differente solubilità in un liquido che si trova all'interno della colonna capillare. Un gas inerte di trasporto fluisce con continuità attraverso il canale della colonna capillare. Aprendo un'apposita valvola, si invia nella colonna una frazione del gas da analizzare che viene trasportato dal gas inerte. Durante il passaggio nella colonna i gas che costituiscono il campione vengono ripetutamente adsorbiti e desorbiti dal sottile strato liquido: ogni gas viene individuato dal tempo di ritenzione nella colonna. Raggiunta l'estremità della colonna, attraverso un foro il flusso passa in un canale situato nella parte posteriore del wafer su cui è montato un rivelatore di conducibilità termica. I gas da analizzare hanno conducibilità termica inferiore a quella dell'elio di trasporto e producono picchi di tensione all'uscita del rivelatore. Il volume di ogni gas è determinato dall'area dei picchi corrispondenti.

Successivamente vengono usate fasi di fotolitografia e di incisione per definire una figura di aperture quadrate nell'ossido sulla superficie inferiore del wafer. Le aperture sono centrate al di sotto dei punti non drogati sulla superficie superiore. Il wafer viene poi inciso con un agente anisotropo dipendente dal drogaggio, la cui azione si arresta quando perviene allo strato fortemente drogato con boro. Tuttavia esso continua ad agire attraverso i punti non drogati con la creazione degli orifizi degli ugelli. Il diametro dei singoli orifizi, lo spessore della membrana circostante e la posizione relativa degli orifizi stessi possono essere controllati con precisione micrometrica poiché non definiti né dallo spessore del wafer né dalla durata della fase di incisione, ma dalle fasi fotolitografica e di diffusione del boro.

Il dispositivo micromeccanico più complesso finora realizzato è un gascromatografo grande come una scatola di fiammiferi da cucina. Due degli autori (Terry e Angell) hanno costruito un prototipo di questo dispositivo nel 1975; il relativo prodotto commerciale è in fase di sviluppo presso la Microsensor Technology, Inc.

La gascromatografia è una tecnica analitica per separare, identificare e misurare la quantità di ogni gas presente in una miscela. Un campione della miscela viene iniettato attraverso una valvola in una lunga colonna capillare, in cui è spinto da un gas inerte di trasporto, che in molti casi è elio. Le pareti della colonna capillare sono ricoperte da un sottile strato di un materiale (per lo più olio di silicone o un polimero) nel quale i vari gas hanno diversi gradi di solubilità. I gas componenti la miscela, nel passaggio attraverso la colonna, sono ripetutamente adsorbiti e desorbiti. Poiché il tempo di adsorbimento di un gas componente dipende dalla sua solubilità, ogni gas viaggia attraverso la colonna a velocità diverse e ne fuoriesce a tempi definiti. Il flusso di uscita viene fatto transitare su un rivelatore che ne misura una proprietà caratteristica, per esempio la conducibilità termica. Il segnale di uscita generato dal rivelatore si pre-

senta sotto forma di una serie di picchi a campana corrispondenti ai gas da analizzare, separati da zone piatte corrispondenti al gas inerte di trasporto. Molti gascromatografi moderni sono associati a un microelaboratore che identifica ciascun gas confrontandone il tempo di ritenzione con tempi campione e misura la quantità di ogni gas componente calcolando l'area delimitata dal picco corrispondente.

Un gascromatografo miniaturizzato è costituito, come un apparecchio convenzionale, da una colonna capillare, da una valvola per l'iniezione del campione e da un rivelatore della conducibilità termica. Tutte le parti del gascromatografo miniaturizzato trovano posto su un solo wafer del diametro di cinque centimetri: il volume della sua colonna capillare è molto inferiore a quello del corrispondente componente dello strumento convenzionale. Un cromatografo funziona in modo corretto solo se il volume del campione di gas iniettato è molto inferiore al volume della colonna; era quindi necessario progettare una valvola miniaturizzata per l'iniezione del gas da accoppiare alla colonna capillare in miniatura. Poiché occorre minimizzare anche il volume compreso fra la valvola e la colonna e quello fra la colonna e il rivelatore, valvola e colonna sono ricavati sullo stesso wafer, mentre il rivelatore è un chip di silicio montato sul wafer stesso.

La colonna capillare è un solco lungo 1,5 metri avvolto a spirale in modo da trovare posto sul wafer, mentre una piastra in vetro fissata al wafer forma la superficie superiore di chiusura della colonna. All'estremità di ingresso della colonna è ricavato nel wafer un foro, attraverso il quale viene introdotto l'elio. A poca distanza un altro foro mette in comunicazione la colonna con la valvola situata sulla superficie posteriore del wafer. Il gas da analizzare entra in un canale separato attraverso un ulteriore foro e fluisce verso la valvola, che lo immette nella colonna capillare.

La sede della valvola sulla faccia posteriore del wafer è costituita da un anello di tenuta in silicio, che circonda entrambi gli orifizi di ingresso e di uscita, e da una sede anulare in silicio che circoscrive il solo orifizio di uscita verso la colonna capillare. Contro l'anello di tenuta è fissato un diaframma di nichel e Teflon. In condizioni di riposo il nucleo di un solenoide è premuto da una molla contro il diaframma, forzandolo sull'anello di tenuta. Quando il solenoide viene eccitato il nucleo arretra e il diaframma si rilascia, consentendo al gas di fluire dall'orifizio di immissione a quello di uscita oltre l'anello di tenuta. Il volume morto effettivo della valvola è costituito dal volume dell'orifizio della colonna capillare, che è pari a quattro nanolitri.

All'estremità di uscita della colonna capillare un altro foro comunica con un canale ricavato per incisione nella superficie posteriore del wafer. Il chip sul quale è realizzato il rivelatore di conducibilità termica è fissato al wafer in posizione capovolta. Il rivelatore è un resistore metallico a film sottile disposto su una membrana termicamente isolante in vetro Pyrex al centro del chip. Nel resistore si fa passare una corrente elettrica costante e si controlla la sua resistenza. I gas da analizzare hanno una conducibilità termica inferiore a quella del gas di trasporto e sottraggono meno calore del resistore, con un conseguente aumento della sua resistenza e la creazione di picchi di tensione. L'ampiezza di ogni picco di tensione è proporzionale alla quantità del gas corrispondente presente nella miscela.

La valvola di iniezione del campione per il gascromatografo miniaturizzato viene fabbricata sullo stesso wafer della colonna capillare. Due fasi di incisione con un agente isotropo definiscono l'anello di tenuta e la sede anulare della valvola, mentre un'incisione con agente anisotropo definisce i fori di alimentazione nel vano della valvola. La prima incisione isotropa crea un anello di tenuta di silicio con bordo affilato sopraelevato di 10 micrometri rispetto al fondo di una cavità. Nella seconda fase si crea nel fondo del primo vano un vano concentrico più piccolo con un altro anello, la sede della valvola, a sua volta sopraelevato di 10 micrometri.

Nella terza fase si formano i fori di alimentazione con un attacco anisotropo che parte dalla superficie posteriore del wafer attraverso aperture quadrate realizzate nell'ossido. Il foro nella sede della valvola si apre nella colonna capillare, mentre gli altri due fori comunicano con i canali del gas da analizzare. Un diaframma in metallo e Teflon e il nucleo di un solenoide completano la valvola. A riposo il diaframma è premuto da una molla contro la sede della valvola, impedendo al gas di fluire nella colonna. Quando il solenoide entra in azione, il diaframma viene rilasciato consentendo al gas sotto pressione di fluire oltre la sede della valvola.

Il vano della valvola, mostrato in questa microfotografia elettronica a scansione, ha un diametro di quattro millimetri e circonda due anelli concentrici di silicio: l'anello di tenuta (all'esterno) e la sede della valvola (all'interno). Le pareti degli anelli sono spesse 50 micrometri. L'orifizio al centro porta alla colonna capillare, mentre gli altri due comunicano con i canali di immissione e di scarico del gas.

La microlavorazione del wafer per gascromatografo ha inizio con l'incisione della sede della valvola. Il vano della valvola, l'anello di tenuta e la sede anulare vengono formati con agenti isotropi attraverso aperture circolari concentriche nello strato di ossido sulla faccia posteriore del wafer. Nella fase successiva si realizzano sia i fori di alimentazione nella valvola e all'estremità della colonna capillare sia i canali per i gas da analizzare tramite attacco da parte di un agente anisotropo attraverso aperture quadrate ricavate nell'ossido della faccia anteriore del wafer. Infine si ottengono la colonna capillare e i canali del gas di trasporto mediante attacco con un agente isotropo attraverso una figura di ossido sulla faccia anteriore.

A questo punto della lavorazione la colonna capillare è un solco a spirale profondo 40 micrometri e largo circa 200 micrometri in superficie. La spirale aperta viene trasformata in un canale chiuso mediante l'asportazione dell'ossido dalla superficie e l'applicazione sul wafer di una piastra in vetro Pyrex con una tecnica, denominata fissaggio anodico, che non richiede la presenza di alcuno strato liquido intermedio, come colla o lega per saldatura. È essenziale usare un metodo di questo tipo, poiché la colonna capillare potrebbe rimanere ostruita da particelle di colla o di lega per saldatura.

Il fissaggio anodico, denominato talvolta anche fissaggio elettrostatico o processo Mallory, fu sviluppato da George Wallie e Daniel I. Pomerantz della P. R. Mallory & Co. La piastra di vetro e il wafer di silicio vengono posti a contatto e riscaldati a circa 400 gradi centigradi. Alla superficie esterna del vetro viene applicata una elevata tensione negativa, mentre il silicio è posto elettricamente a massa. Il vetro contiene una piccola quantità di sodio che a quella temperatura ionizza e diventa mobile. Gli ioni sodio positivi migrano verso l'elettrodo negativo, lasciando dietro di loro nel vetro cariche negative legate in prossimità dell'interfaccia fra il vetro stesso e il silicio. Di conseguenza la tensione fra vetro e silicio è pressoché uguale alla differenza di potenziale applicata: l'attrazione elettrostatica forza il vetro e il silicio a entrare in stretto contatto e a fondersi, creando una tenuta ermetica.

Il rilevatore di conducibilità termica sviluppato da John H. Jerman di Stanford per il gascromatografo miniaturizzato è un chip costruito su un wafer separato. Questo viene ossidato e ricoperto di uno strato di vetro Pyrex depositato con una tecnica denominata *sputtering* (letteralmente «spruzzamento»). Sulla parete superiore del vetro vengono depositati resistori metallici a film sottile. Nella fase successiva l'incisione con un agente anisotropo attraverso aperture quadrate praticate nell'ossido della faccia posteriore del wafer asporta tutto lo spessore di silicio, lasciando una membrana isolante di vetro sotto ogni serie di resistori. Il wafer è poi segato in chip, ai quali si collegano dei conduttori; la faccia anteriore del chip è infine fissata sul canale del gas all'estremità della colonna capillare.

Grazie al volume estremamente ridotto della colonna capillare, il gascromatografo miniaturizzato richiede quantitativi di gas di trasporto molto inferiori a quelli richiesti da un gascromatografo convenzionale. Questa caratteristica, unita alle piccole dimensioni dell'unità stessa, permette di costruire gascromatografi portatili. In effetti, uno strumento portatile costituito da cinque gascromatografi miniaturizzati e da un microcalcolatore è in corso di sviluppo presso la Microsensor

Il rivelatore di conducibilità termica del gascromatografo è un resistore realizzato mediante incisione di un sottile film metallico depositato su un wafer che è stato rivestito in precedenza con vetro. Dopo la formazione del resistore, il wafer è inciso anisotropicamente dalla parte posteriore e rimane solo una sottile membrana di vetro sotto il resistore. Poiché il vetro è un cattivo conduttore di calore (al contrario del silicio che è un buon conduttore), la membrana isola termicamente il resistore. Una variazione di conducibilità termica nel flusso di gas che scorre sopra il resistore fa sì che ne venga modificata la temperatura e quindi la resistenza elettrica.

Technology. Se le cinque colonne capillari sono rivestite con materiali differenti, si può avere uno strumento sensibile a circa 100 gas. Un tale strumento può controllare la qualità dell'aria o misurare il contenuto energetico del gas in un gasdotto.

La microlavorazione è stata usata recentemente per realizzare, in un wafer di silicio, dissipatori di calore molto efficienti. Tali dissipatori sono impiegati per migliorare le prestazioni di circuiti integrati ad altissima velocità. In generale l'energia dissipata in un dispositivo elettronico aumenta con la sua velocità e quindi il calore prodotto da circuiti più veloci può portare alla loro distruzione, se non sono opportunamente raffreddati. Anche con l'adozione di particolari accorgimenti di incapsulamento, il limite superiore di dissipazione dei circuiti ad alta velocità è di circa 20 watt per centimetro quadrato di chip.

L'illustrazione mostra un pressosensore in silicio a metà del processo di fabbricazione. Il quadrato incassato al centro della microfotografia è un sottile diaframma, sensibile alla pressione, con lato di 0,5 millimetri. Gli elementi di colore violetto, dalla forma a serpentina, presenti sul diaframma sono resistori ottenuti per diffusione di boro nel wafer. Le regioni in verde, che fungono da conduttori fra i resistori, hanno una maggiore concentrazione di boro. I rettangoli neri ai bordi del sensore sono aree di saldatura incomplete, costituite da profondi solchi a *V* che verranno dorati e a cui saranno fissati conduttori di collegamento. Quando il diaframma si flette, la resistenza delle zone violette cambia sotto l'azione della sollecitazione meccanica, un effetto detto piezoresistivo. I resistori sono disposti in modo che, quando la resistenza di due di essi aumenta, quella degli altri due diminuisce; la configurazione a ponte dei resistori ha una sensibilità alle variazioni di pressione doppia di quella del resistore singolo.

In questa sezione di un pressosensore lo spessore del diaframma è stato esagerato per riuscire a evidenziarne i dettagli. Il diaframma è stato realizzato con una incisione anisotropa poco profonda nella parte superiore del wafer e con una incisione profonda nella parte inferiore. Quando il sensore è completo, la superficie superiore viene fissata a una piastra di vetro, in modo che si formi una cavità sigillata, mentre il fondo del diaframma viene esposto alla pressione da misurare. Essendo nota la pressione nella cavità, l'uscita è in grado di misurare la pressione assoluta.

David B. Tuckerman e Fabian Pease dello Stanford Integrated Circuits Laboratory hanno sviluppato un dissipatore che consente densità di potenza molto superiori. La loro idea consiste nel fabbricare dispositivi elettronici sulla superficie anteriore di un wafer nella cui parte posteriore sono ricavati solchi rettangolari molto ravvicinati l'uno all'altro. Attraverso i canali così formati viene fatto circolare un liquido refrigerante. Il silicio possiede una conducibilità termica elevatissima e Tuckerman e Pease hanno realizzato dispositivi ai quali lo spessore delle pareti, la larghezza e la profondità dei canali impartiscono il massimo rendimento di raffreddamento.

I solchi rettangolari sono ottenuti con un procedimento assai simile a quello descritto in precedenza per i solchi a *V*. Un agente anisotropo crea una serie di solchi a *V* molto ravvicinati partendo da una figura di linee parallele sull'ossido se il wafer è del tipo <100>, come quelli citati finora. Se invece il wafer è del tipo <110>, un agente anisotropo crea, partendo dalla medesima figura sull'ossido, solchi con pareti laterali verticali e sezione trasversale all'incirca rettangolare. I solchi rettangolari sono preferibili per i dissipatori perché possono essere più ravvicinati di quelli a *V*, ed esporre quindi una maggiore superficie al liquido di raffreddamento circolante. Dopo la formazione dei solchi il wafer viene fissato anodicamente al vetro per ottenere dei microcanali chiusi. Un dissipatore con canali incisi profondi 300 micrometri e larghi 50 micrometri, separati da alette di silicio larghe 100 micrometri, e nei quali scorre acqua per circolazione forzata, possono consentire di dissipare potenze di oltre 1000 watt per centimetro quadrato.

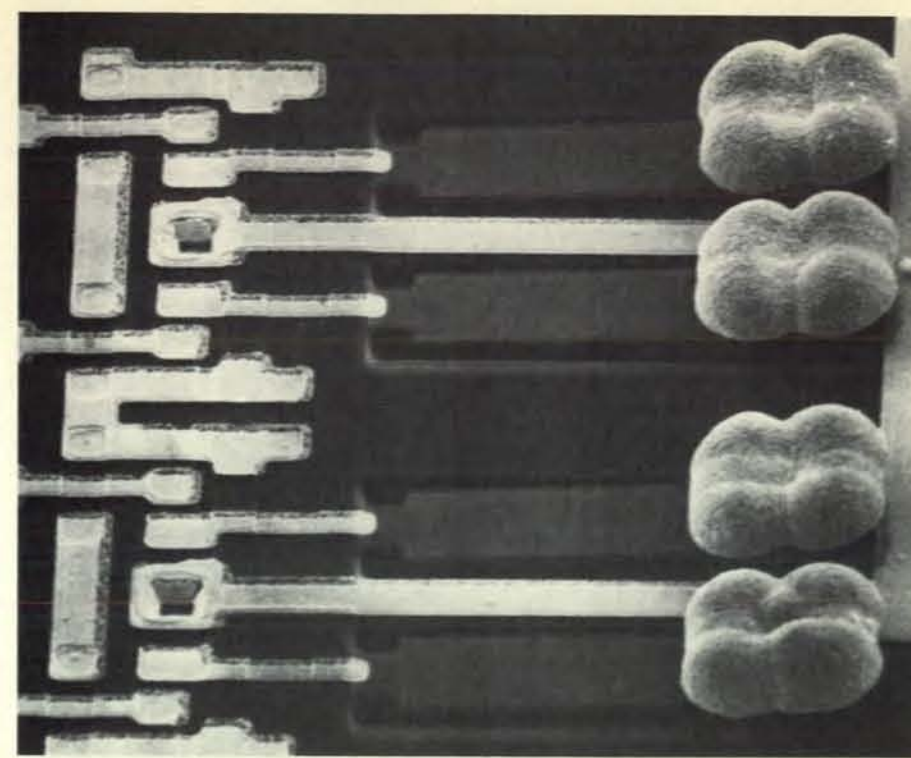
È possibile costruire dissipatori del tipo illustrato solo in silicio e solo su wafer con un particolare orientamento cristallino. Tuckerman e Pease hanno studiato anche la possibilità di creare dissipatori termici segando meccanicamente il materiale anziché inciderlo. Una sega circolare con un piccolo bordo metallico ricoperto di polvere di diamante può tagliare nella faccia posteriore di un wafer canali larghi da 50 a 200 micrometri. Anche se le dimensioni dei canali ottenuti in tal modo non sono controllabili con la stessa precisione consentita dall'incisione, la lavorazione meccanica permette di preparare dissipatori in wafer di silicio di qualunque orientamento o in wafer di altro materiale, come il semiconduttore composto da arseniuro di gallio e da zaffiro, usato quale substrato per alcuni tipi di circuiti integrati. I wafer realizzati in tale materiale possono inoltre essere tagliati secondo uno schema incrociato, di modo che i circuiti sono supportati da microcolonne anziché da microalette e i dissipatori sono meno soggetti a occlusioni da parte di particelle nel liquido di raffreddamento.

Pressosensori e accelerometri sono due esempi di trasduttori integrati in silicio. Il più piccolo accelerometro finora costruito è un dispositivo a sbalzo con amplificatore contenuto sullo stesso chip. L'accelerometro, progettato da Petersen, Anne C. Shartel e Norman F. Raley del San Jose Research Laboratory della IBM, è costituito da un elemento a sbalzo in ossido sospeso su una cavità poco profonda realizzata con la tecnica dello strato di arresto drogato con boro. Sulla superficie superiore dell'elemento viene depositato uno strato di metallo che, insieme con il silicio della cavità, agisce come un condensatore variabile ad aria. Sulla estremità libera dell'elemento viene formato per placcatura un deposito di oro. Se il chip di silicio viene spostato improvvisamente, l'inerzia della massa di oro flette l'elemento, modificando la geometria del condensatore e quindi la sua capacità.

L'uscita del sensore è una tensione proporzionale all'accelerazione. Lo strumento ha una sensibilità di due millivolt per *g* (l'unità di accelerazione pari all'accelerazione di gravità in corrispondenza del livello del mare). Bisogna tener presente che l'amplificatore è parte essenziale del dispositivo: la variazione di capacità da rilevare è piccolissima e, se il segnale non venisse amplificato sullo stesso chip, rimarrebbe del tutto mascherata da ben maggiori variazioni di capacità introdotte da qualsiasi conduttore connesso al sensore. Tale osservazione è valida in generale per tutti i piccoli sensori capacitivi: la trasmissione deve essere preceduta da qualche trattamento del segnale.

I sensori in silicio di maggiore diffusione sono i pressosensori, alcuni dei quali comprendono dispositivi elettronici integrati. Esistono due tipi fondamentali di pressosensori in silicio: piezoresistivi e capacitivi. Il resistore elettrico sul sottile diaframma flessibile di un sensore piezoresistivo cambia valore di resistenza quando il diaframma cambia assetto. In un sensore capacitivo il diaframma flessibile costituisce una delle armature di un condensatore variabile ad aria.

Nella maggior parte dei casi il diaframma di un pressosensore è ottenuto per incisione con agente anisotropo in un wafer di silicio <100>. Per ottenere diaframmi con spessori compresi fra 10 e 50 micrometri si può ricorrere semplicemente a una incisione a tempo, mentre per diaframmi più sottili è necessario ricorrere a strati di arresto dell'incisione.

I resistori di un sensore piezoresistivo si formano drogando determinate zone del diaframma con uno dei due tipi di droganti disponibili. Gli elementi a destra del silicio nella tavola periodica, come il fosforo, hanno cinque elettroni di valenza, mentre il silicio ne ha quattro: l'elettrone

Gli accelerometri capacitivi sono elementi in biossido di silicio realizzati a sbalzo su una cavità poco profonda nel silicio. In questa microfotografia realizzata al microscopio elettronico a scansione si vedono quattro accelerometri. Le piccole masse a destra sono d'oro e aumentano l'inerzia di ciascun elemento e quindi la sensibilità del dispositivo. Uno strato metallico (al di sopra dell'elemento), uno strato di arresto dell'incisione (realizzato in boro ed elettricamente conduttore, posto sul fondo della cavità) e il cuscinetto d'aria tra di essi formano un condensatore variabile. Sottoposto ad accelerazione l'elemento si flette, variando la capacità del condensatore. Il circuito a sinistra amplifica il segnale; l'uscita del sensore è una tensione proporzionale all'accelerazione.

in più può diventare portatore di corrente elettrica. Gli elementi a sinistra del silicio nella tavola periodica, come il boro, hanno tre elettroni di valenza: anche l'assenza di un elettrone (una «buca») può agire come portatore di corrente, sebbene le buche siano meno mobili degli elettroni. Non può esistere flusso di corrente fra una regione drogata con fosforo e una drogata con boro a causa della diversa natura dei portatori di carica. Quindi se il silicio di cui è costituito il wafer è stato leggermente drogato con fosforo, una più elevata concentrazione di boro può essere diffusa in zone prescelte del wafer per ottenere resistori elettricamente isolati dal resto del diaframma.

L'effetto piezoresistivo si manifesta come una variazione di resistenza elettrica prodotta da una sollecitazione meccanica: l'ampiezza dell'effetto dipende dall'orientamento del resistore. Per fortunata coincidenza i resistori ottenuti per drogaggio con boro in un wafer <100> offrono la massima sensibilità alla sollecitazione quando sono allineati con i bordi <110> di una cavità ottenuta con un agente anisotropo.

Gli autori hanno sviluppato un pressosensore piezoresistivo con quattro resistori disposti su un diaframma quadrato. I resistori sono situati in prossimità dei bordi del diaframma (le zone di massima sollecitazione) e sono allineati con le direzioni <110> (lungo le quali è maggiore la risposta alla sollecitazione). I resistori sono collegati secondo la configurazione comunemente nota come ponte di Wheatstone. Quando il diaframma si deforma, la resistenza di due resistori aumenta e quella dei due rimanenti diminuisce della stessa quantità: ne consegue che il ponte ha una sensibilità alle variazioni di pressione doppia di quella di un resistore semplice.

Il primo chip sensibile alla pressione e comprendente circuiti attivi (cioè elementi circuitali che amplificano il segnale) fu progettato da John M. Borky e Kensall Wise dell'Università del Michigan. Il circuito converte la tensione di uscita del ponte di Wheatstone in un segnale a modulazione di frequenza: in altre parole il chip converte le variazioni di pressione in variazioni di frequenza. Diversi sensori di questo tipo, cui siano assegnate diverse gamme di frequenza, possono ripartirsi una linea di trasmissione nello stesso modo in cui diverse stazioni radio a modulazione di frequenza si ripartiscono la banda a radiofrequenza loro assegnata.

I pressosensori capacitivi possiedono una sensibilità intrinseca alle variazioni di pressione maggiore dei sensori piezoresistivi, ma, come è stato notato in precedenza, il segnale prodotto da un sensore capacitivo può essere facilmente mascherato dal rumore. I circuiti di elaborazione del segnale, che per i sensori piezoresistivi consentono di migliorarne le prestazioni,

sono essenziali per un sensore capacitivo. Craig S. Sander, James W. Knutti e James D. Meindl dello Stanford Integrated Circuits Laboratory hanno progettato il primo pressosensore capacitivo con circuito attivo sullo stesso chip. Il diaframma di questo sensore, incassato di circa cinque micrometri nel wafer di silicio, costituisce una delle armature di un condensatore; l'altra armatura è un film metallico su una lastra di vetro fissata anodicamente al wafer.

Nella cavità fra diaframma e piastra di vetro rivestita di metallo rimane sigillata una certa quantità di gas. L'altra faccia del diaframma è esposta alla pressione da misurare: quando questa cambia e il diaframma si deforma, varia lo spessore del gas isolante posto fra metallo e silicio. Il circuito attivo sul chip controlla la risultante variazione di capacità, erogando una serie di impulsi di tensione. Il tempo fra due impulsi successivi è proporzionale alla pressione misurata. Questo segnale, codificato a impulsi nel tempo, è meno facilmente soggetto al rumore di un segnale di tensione semplice non codificato, in modo analogo a quanto si verifica per il segnale a modulazione di frequenza ottenuto da un pressosensore integrato piezoresistivo.

I sensori integrati possono soddisfare i requisiti di una maggiore varietà di applicazioni dei sensori discreti, poiché i loro segnali di uscita sono relativamente insensibili al rumore e alla dispersione. I sensori integrati possono quindi essere montati in ambienti difficili, dove saranno sottoposti a sollecitazioni severe, come sotto il cofano di un'automobile, o in situazioni nelle quali il dispositivo sensibile deve essere separato fisicamente dall'elettronica di elaborazione del segnale, come nei controlli biomedici. Molti dei sensori oggi in corso di sviluppo sono previsti per tali difficili applicazioni: due di noi (Barth e Angell) stanno per esempio lavorando su sensori da impiantare in tessuti viventi.

Sono in fase di sviluppo anche chip che comprendono vari tipi di sensori, come un sensore per la temperatura e uno per la pressione, oppure più esemplari dello stesso sensore (per aumentare l'affidabilità) e anche altri chip che dispongono di qualche mezzo di taratura dei sensori. In futuro molti sensori integrati saranno in grado di convertire segnali analogici in segnali digitali e potranno così essere direttamente collegati a elettroniche digitali. In generale il maggiore interrogativo attuale è la determinazione di quanti circuiti sia possibile ed economicamente conveniente predisporre su un chip sensore.

Mentre viene accordata una grande attenzione ai sensori, continua lo sviluppo dei dispositivi esclusivamente meccanici in silicio. Le tecniche di incisione del silicio sono tuttora in fase di perfezionamento e dai laboratori usciranno nei prossimi anni nuove strutture micromeccaniche. I dispositivi micromeccanici in silicio potrebbero essere un giorno comuni quasi quanto lo sono oggi quelli elettronici.

La struttura dei quark e dei leptoni

Non si esclude che quark e leptoni, considerati particelle elementari, siano formati da entità ancora più piccole confinate entro un volume di dimensioni inferiori a un millesimo delle dimensioni del protone

di Haim Harari

Negli ultimi 100 anni la ricerca dei costituenti ultimi della materia ha attraversato quattro strati di struttura. Si è visto che tutta la materia è costituita di atomi e che l'atomo stesso ha un denso nucleo circondato da una nube di elettroni. Il nucleo è stato a sua volta diviso nei protoni e neutroni componenti e più recentemente è apparso chiaro che anche il protone e il neutrone sono particelle composte, formate da entità più piccole, cui si è dato il nome di quark. Che cosa viene dopo? È del tutto probabile che la progressione orbita dentro orbita sia alla fine giunta al termine e che non sia possibile suddividere ulteriormente i quark. Anche i leptoni, la classe di particelle che comprende l'elettrone, potrebbero essere elementari e indivisibili. Alcuni fisici non sono però sicuri che si sia portato alla luce il «nocciolo» più interno della materia. Essi hanno incominciato a domandarsi perché anche i quark e i leptoni non debbano avere una qualche composizione interna.

L'impulso determinante per la ricerca di un ulteriore strato di struttura è venuto dalla convinzione (o forse dal pregiudizio) che debbano esistere soltanto pochi mattoni fondamentali della materia. L'economia di mezzi è stata a lungo un principio-guida della fisica ed è stata utile finora. L'elenco dei costituenti fondamentali della materia si allungò dapprima incredibilmente verso la fine del XIX secolo, quando il numero degli elementi chimici, e quindi il numero delle specie atomiche, si avvicinò a 100. La risoluzione della struttura atomica risolse il problema e verso il 1935 il numero di particelle elementari era fermo a quattro: il protone, il neutrone, l'elettrone e il neutrino. Questa visione parsimoniosa del mondo venne meno negli anni cinquanta e sessanta; risultò infatti che il protone e il neutrone sono i rappresentanti di una grandissima famiglia di particelle, la famiglia chiamata oggi degli adroni. Verso la metà degli anni sessanta il numero di forme fondamentali della materia era ancora di circa 100. Questa volta fu il modello a quark a portare chiarezza. Nella formulazione iniziale del modello tutti gli adroni potevano essere spiegati come combinazioni di tre soli tipi di quark.

Oggi è la proliferazione degli stessi quark e leptoni che sta cominciando a destare interesse sulla possibilità di uno schema più semplice. Mentre il modello originale aveva tre quark, oggi si pensa che ne esistano almeno 18, oltre a sei leptoni e a una dozzina di altre particelle che agiscono come portatori di forze. Tre dozzine di unità fondamentali della materia sono troppe per il gusto di alcuni fisici e non esiste la certezza che non vengano scoperti altri quark e leptoni. Il modo forse più diretto di ridurre il caos è postulare un livello di organizzazione ancor più profondo. Tutti i quark e i leptoni sarebbero in tal caso oggetti composti, allo stesso modo degli atomi e degli adroni, e dovrebbero la loro varietà al numero di modi in cui si possono associare alcuni costituenti più piccoli. La diversità della natura attualmente osservata non sarebbe intrinseca, ma combinatoria.

Andrebbe notato che finora non esistono prove che i quark e i leptoni abbiano una struttura interna di alcun genere. Nel caso dei leptoni gli esperimenti sono giunti fino a 10^{-16} centimetri e non hanno scoperto nulla che contraddica l'ipotesi che i leptoni siano puntiformi e privi di struttura. Per quanto riguarda i quark, non è stato finora possibile esaminare un quark isolato, né tantomeno distinguere qualsiasi possibilità di strutture interne. Anche come concezione strettamente teorica, l'idea delle subparticelle incontra difficoltà: nessuno è stato in grado di presentare una descrizione coerente del possibile moto delle subparticelle all'interno di un quark o di un leptone e delle possibili interazioni reciproche. Esse dovrebbero essere quasi incredibilmente piccole: se un atomo venisse ingrandito alle dimensioni della Terra, i suoi costituenti più interni non potrebbero essere più grandi di un acino d'uva. Ciononostante, i modelli della sottostruttura di quark e leptoni fanno appello al senso estetico e all'immaginazione: essi suggeriscono un modo per costruire con poche parti semplici un mondo complesso.

Qualsiasi teoria delle particelle elementari della materia deve anche prendere in considerazione le forze che agiscono tra esse e le leggi della natura che governano tali forze. Si guadagnerebbe poco semplificando la gamma delle particelle, se in questo modo si aumentasse il numero delle forze e delle leggi. Come accade, in tutta la storia della fisica, vi è stato un sottile interscambio tra l'elenco delle particelle e l'elenco delle forze.

Nell'Ottocento si pensava che le forze fondamentali fossero quattro: la gravitazione, l'elettricità, il magnetismo e la forza a breve raggio d'azione tra le molecole che è responsabile della coesione della materia. Una serie di importanti scoperte sperimentali e teoriche condusse poi a riconoscere che in realtà l'elettricità e il magnetismo sono due manifestazioni della stessa forza fondamentale, che venne ben presto chiamata «elettromagnetismo». La scoperta della struttura atomica introdusse un'ulteriore revisione. Benché un atomo nel suo complesso sia elettricamente neutro, i suoi costituenti sono carichi e la forza molecolare a breve raggio d'azione finì con l'essere spiegata come un complesso effetto residuo delle interazioni elettromagnetiche tra nuclei positivi ed elettroni negativi. Quando due atomi neutri sono molto lontani, tra essi non si manifestano in pratica forze elettromagnetiche. Quando, invece, sono vicini, i costituenti carichi di un atomo sono in grado di «vedere» e di influenzare le cariche più interne dell'altro, causando varie attrazioni e repulsioni a breve raggio.

In seguito a queste scoperte, la fisica restò con due sole forze fondamentali. L'unificazione dell'elettricità con il magnetismo aveva ridotto il numero di un'unità e l'interazione molecolare era stata retrocessa dal ruolo di forza fondamentale a quello di forza derivata. Le due restanti forze fondamentali, la gravitazione e l'elettromagnetismo, erano entrambe a lungo raggio d'azione. L'esplorazione della struttura nucleare introdusse tuttavia ben presto due nuove forze a breve raggio d'azione: la forza forte che tiene assieme i protoni e i neutroni nel nucleo e la forza debole che media certe trasformazioni di una particella in un'altra, come nel decadimento beta di un nucleo radioattivo. In tal modo le forze diventarono di nuovo quattro.

Lo sviluppo del modello a quark e la relativa teoria delle interazioni tra quark costituirono l'occasione successiva per rivedere l'elenco delle forze. Si pensa che i quark in un protone o in un neutrone siano tenuti assieme da un nuovo tipo di forza fondamentale a lungo raggio d'azione, chiamata forza di colore, la quale agisce sui quark dal momento che essi recano un nuovo tipo di carica chiamata colore. (Né la forza né la carica hanno nulla a che vedere con i colori ordinari.) Allo stesso modo in cui un atomo è formato da costituenti elettricamente carichi pur essendo di per sé neutro, così un protone o un neutrone sono formati da quark colorati pur essendo di per sé incolori. Quando due protoni incolori sono molto lontani, non vi sono in pratica forze di colore tra essi, ma quando sono vicini i quark colorati di un protone «vedono» le cariche di colore dell'altro protone. Le attrazioni e le repulsioni a breve raggio d'azione che ne risultano sono state identificate negli effetti della forza forte. In altre parole, allo stesso modo in cui la forza molecolare a breve raggio d'azione divenne un residuo della forza elettromagnetica a lungo raggio d'azione, la forza forte a breve raggio d'azione è diventata un residuo della forza di colore a lungo raggio d'azione.

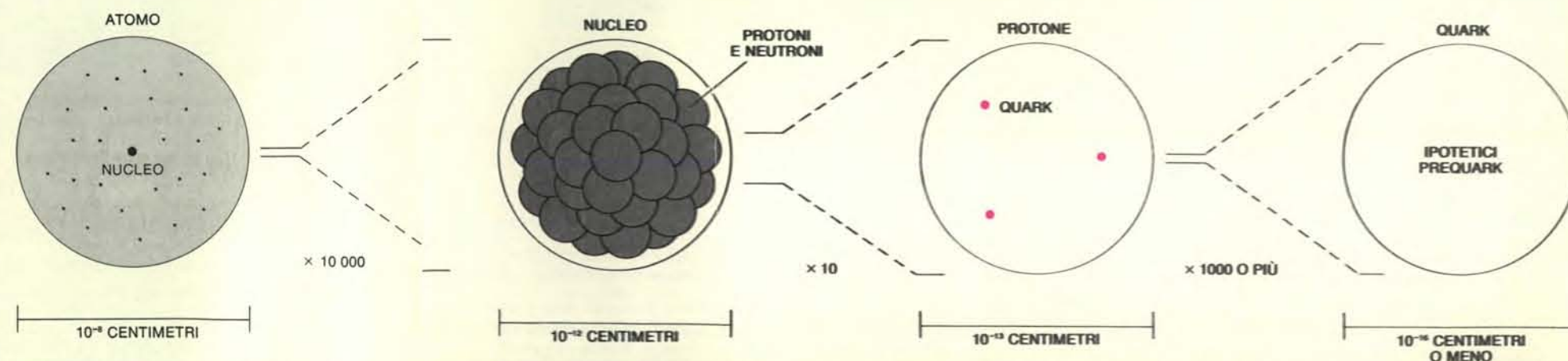
Si può aggiungere un altro capitolo a questa breve storia delle forze della natura. È stata scoperta una profonda e suggestiva connessione tra l'elettromagnetismo e la forza debole, portandole quasi al punto di una completa unificazione. Esse sono chiaramente correlate, ma la correlazione non è così stretta come nel caso dell'elettricità e del magnetismo e, quindi, esse devono ancora essere considerate come forze separate. L'attuale elenco delle forze fondamentali presenta perciò ancora quattro componenti: le forze gravitazionale, elettromagnetica e di colore, tutte a lungo raggio d'azione, e la forza debole a breve raggio d'azione. Nei limiti delle conoscenze attuali tutti i fenomeni naturali si possono spiegare mediante queste forze e i loro effetti residui.

L'evoluzione delle idee sulle particelle e sulle forze sono chiaramente interdipendenti. Man mano che si scoprono nuove particelle fondamentali, si vede che le vecchie sono oggetti composti. Man mano che si scoprono nuove forze, le vecchie vengono unificate o ridotte a stati residui. Gli elenchi delle particelle e delle forze vengono revisionati di tanto in tanto man mano che la materia viene esplorata a una scala più piccola e che le conoscenze teoriche progrediscono. Qualsiasi variazione in un elenco conduce inevitabilmente a una modificazione dell'altro. Le recenti ipotesi sulla struttura dei quark e dei leptoni non fanno eccezione; anch'esse richiedono variazioni nella definizione delle forze. Resta da vedere se le modificazioni rappresentano una semplificazione.

Delle quattro forze fondamentali definite, la gravitazione deve essere inserita in una categoria a sé. Essa è troppo debole perfino per essere rilevata nelle interazioni di particelle singole e non si spiega in termini di eventi microscopici. Per le altre tre forze sono state elaborate teorie di successo oggi ampiamente accettate. Le tre teorie sono distinte, ma reciprocamente coerenti; prese insieme, esse costituiscono un modello esauriente delle particelle elementari e delle loro interazioni, al quale mi riferirò come modello standard.

Nel modello standard i costituenti indivisibili della materia sono i quark e i leptoni. È opportuno trattare per primi i leptoni. Ne esistono sei: l'elettrone e il suo compagno il neutrino elettronico, il muone e il neutrino muonico, la tau e il neutrino tauonico. L'elettrone, il muone e la tau hanno una carica elettrica −1, mentre i tre neutrini sono elettricamente neutri.

Esistono anche sei tipi fondamentali di quark, ai quali sono stati dati i nomi di su *(up)*, giù *(down)*, incantato *(charmed)*, strano *(strange)*, alto *(top)* e basso *(bottom)* o *u,d,c,s,t* e *b*. (Il quark alto non è ancora stato rivelato sperimentalmente e nemmeno il neutrino tauonico, ma pochi teorici dubitano della loro esistenza.) I quark *u,c*, e *t* hanno una carica elettrica +2/3, i quark *d,s*, e *b* una carica elettrica −1/3. Inoltre, ogni tipo di quark ha tre possibili colori, che indicherò con rosso, giallo e blu. Così, se si considera come particella a se stante ogni quark colorato, esistono complessivamente 18 varietà di quark. Si osservi che mentre ciascun quark reca sia un colore sia una carica elettrica, nessuno dei leptoni è colorato.

Per ogni particella di questo schema

La gerarchia delle particelle nella struttura della materia presenta attualmente quattro livelli. Tutta la materia è costituita da atomi; l'atomo è formato da un nucleo circondato da elettroni; il nucleo è composto da protoni e neutroni; tutti i protoni e i neutroni sarebbero composti da tre quark. Secondo alcune ipotesi formulate di recente si potrebbe aggiungere un quinto livello: il quark potrebbe essere composto da ipotetici costituenti più piccoli (i prequark). Anche i leptoni, la classe di particelle che comprende l'elettrone, potrebbero essere formati da prequark.

esiste un'antiparticella con la stessa massa, ma con valori opposti della carica elettrica e del colore. L'antiparticella dell'elettrone è il positone, dotato di carica +1. L'antiparticella di un quark rosso *u*, con carica +2/3, è un antiquark antirosso *ū*, con carica −2/3.

La proprietà del colore dei quark è sotto molti aspetti analoga alla carica elettrica, ma, essendovi tre colori possibili, è decisamente più complessa. Le particelle elettricamente cariche si possono riunire per formare un sistema elettricamente neutro in un solo modo: combinando uguali quantità di carica positiva e negativa. Una particella composta incolore si può formare con quark colorati all'incirca allo stesso modo, cioè combinando un quark colorato e un antiquark anticolorato. Tuttavia, nel caso del colore, esiste un secondo modo di formare uno stato neutro: qualsiasi sistema composto con uguali quantità dei tre colori e dei tre anticolori è anch'esso incolore. Per questo motivo un protone formato da un quark rosso, da un quark giallo e da un quark blu non ha alcun colore risultante.

Le forze fondamentali della natura si possono classificare secondo uno schema che si è evoluto insieme all'elenco delle particelle elementari. Qui sono indicate in grigio le forze a lungo raggio d'azione e in colore quelle a breve raggio d'azione. All'inizio del XIX secolo tre forze a lungo raggio d'azione erano considerate fondamentali: quella gravitazionale, quella elettrica e quella magnetica. Anche una forza a breve raggio d'azione, la forza molecolare, responsabile della coesione della materia, era considerata fondamentale. James Clerk Maxwell unificò l'elettricità e il magnetismo e con la scoperta della struttura atomica apparve chiaro che la forza molecolare non è fondamentale, ma è invece un effetto residuo dell'elettromagnetismo. La scoperta del nucleo atomico introdusse due forze a breve raggio d'azione, quella debole e quella forte. Nel modello a quark, però, la forza forte diventa un residuo di una nuova forza a lungo raggio d'azione chiamata forza di colore. È stata inoltre scoperta una profonda relazione tra la forza debole e l'elettromagnetismo, di modo che possono essere considerate parzialmente unificate. La sesta riga del grafico rappresenta le forze della natura interpretate secondo il «modello standard» della fisica delle particelle elementari. Un valido modello della struttura dei quark e dei leptoni potrebbe portare a un'ulteriore revisione. In alcuni modelli, per esempio, vi è una nuova forza a lungo raggio d'azione chiamata «ipercolore» di cui la forza debole è un residuo. In modelli di questo tipo tutte le forze fondamentali della natura sono a lungo raggio d'azione. Tali modelli sono però ancora altamente ipotetici.

Dovremmo menzionare un'altra proprietà dei quark e dei leptoni: ogni particella è dotata di uno spin, o momento angolare intrinseco, uguale alla metà dell'unità fondamentale quantomeccanica di momento angolare. Quando una particella con spin 1/2 si muove in linea retta, la sua rotazione intrinseca può avvenire in senso orario o in senso antiorario, osservando la particella lungo la direzione di moto. Se lo spin è in senso orario, si dice che la particella è destrorsa, perché, ripiegando le dita della mano destra nella stessa direzione dello spin, il pollice indica la direzione del moto. Per una particella con verso di spin opposto, il moto viene descritto dalla regola della mano sinistra e la particella si dice sinistrorsa.

Nel modello standard le tre forze che agiscono sui quark e sui leptoni sono descritte essenzialmente dalla stessa struttura matematica. È nota come teoria di campo a invarianza di gauge o semplicemente come teoria di gauge. Ogni forza viene trasmessa da una particella all'altra da campi portatori i quali, a loro volta, sono contenuti nelle particelle portatrici, o bosoni di gauge.

La teoria di gauge della forza elettromagnetica, chiamata elettrodinamica quantistica o QED, è la prima e la più semplice delle tre teorie. Essa è stata proposta negli anni quaranta da Richard P. Feynman, Julian S. Schwinger e Sin-Itiro Tomonaga. La QED descrive le interazioni di particelle elettricamente cariche, in particolare l'elettrone e il positone. Esiste un tipo di bosone di gauge che media le interazioni; si tratta del fotone, il ben noto quanto della radiazione elettromagnetica, privo di massa e di carica elettrica propria. QED è probabilmente la teoria fisica più accuratamente verificata. Per esempio, essa prevede correttamente il momento magnetico dell'elettrone con almeno 10 cifre significative.

La teoria della forza di colore è stata formulata in analogia con la QED ed è chiamata cromodinamica quantistica, o QCD. Essa è stata messa a punto in un periodo di quasi vent'anni con gli sforzi di numerosi fisici teorici. In QCD le particelle interagiscono in virtù del loro colore anziché della loro carica elettrica. I bosoni di gauge di QCD, responsabili del legame tra i quark all'interno di un adrone, sono chiamati gluoni. Come il fotone, i gluoni sono privi di massa, ma mentre esiste un solo tipo di fotone, esistono otto specie di gluoni. C'è un'altra differenza ancora più importante tra il fotone e il gluone. Pur essendo il fotone il mediatore della forza elettromagnetica, esso non ha carica elettrica e quindi non dà origine ad alcuna forza elettromagnetica propria (o almeno ad alcuna forza di entità significativa). I gluoni, al contrario, non sono incolori. Essi trasmettono la forza di colore tra i quark, ma hanno anche colore proprio e sono sensibili alla forza di colore. Questa proprietà riflessiva, per effetto della quale il portatore della forza agisce su se stesso, rende estremamente difficile una completa analisi matematica della forza di colore.

		QUARK		LEPTONI	
GENERAZIONI	PRIMA	*u* (SU)	*d* (GIÙ)	*e* (ELETTRONE)	ν_e NEUTRINO ELETTRONICO
	SECONDA	*c* (INCANTATO)	*s* (STRANO)	μ (MUONE)	ν_μ NEUTRINO MUONICO
	TERZA	*t* (ALTO)	*b* (BASSO)	τ (TAU)	ν_τ NEUTRINO TAUONICO
CARICHE	ELETTRICA	+2/3	−1/3	−1	0
	DI COLORE	ROSSO GIALLO BLU	ROSSO GIALLO BLU	INCOLORE	INCOLORE
FORZE	DI COLORE	■	■		
	ELETTRO-MAGNETICA	■	■	■	
	DEBOLE	■	■	■	■

Il modello standard delle particelle elementari comprende tre «generazioni» di quark e di leptoni, anche se tutta la materia comune si può costruire con le sole particelle della prima generazione. I quark sono caratterizzati da valori frazionari della carica elettrica e da una proprietà fantasiosamente chiamata colore: ogni tipo di quark può apparire nelle tre versioni rosso, giallo e blu. I leptoni hanno cariche elettriche intere e sono incolori. Le due classi di particelle differiscono anche per la loro risposta alle varie forze. Solo i quark sono soggetti alla forza di colore e quindi possono restare permanentemente confinati all'interno di particelle composte quale il protone.

Una peculiarità che pare riguardare QCD è il fenomeno del confinamento del colore. Si ritiene che la forza di colore intrappoli in qualche modo oggetti colorati (quali i quark e i gluoni) all'interno di oggetti composti che sono immancabilmente incolori (quali i protoni e i neutroni). Le particelle colorate non possono mai sfuggire (pur potendo formare nuove combinazioni incolori). I fisici ritengono che a causa del confinamento del colore non si sono mai visti un quark o un gluone isolati. Devo ammettere che, pur se l'ipotesi del colore è oggi ampiamente accettata, non si è dimostrato che derivi da QCD. Ci possono essere altre sorprese in serbo.

La forza debole è alquanto diversa dalle altre due, ma ciononostante essa può essere descritta da una teoria di gauge dello stesso tipo generale. Negli anni sessanta e all'inizio degli anni settanta la teoria è stata elaborata ed è stata fissata l'importante correlazione tra la forza debole e la forza elettromagnetica, da parte di numerosi ricercatori. Notevoli furono i contributi apportati (in ordine cronologico) da Sheldon Lee Glashow della Harvard University, da Steven Weinberg dell'Università del Texas ad Austin, da Abdus Salam del Centro internazionale di fisica teorica di Trieste e da Gerard 't Hooft dell'Università di Utrecht.

È curioso il fatto che le cariche sulle quali agisce la forza debole siano associate al carattere destrorso o sinistrorso della particella. Sia tra i quark sia tra i leptoni le particelle sinistrorse e quelle destrorse hanno una carica debole, mentre le antiparticelle destrorse e quelle sinistrorse sono neutre rispetto alla forza debole. Ciò che è ancora più strano, la carica debole non si conserva in natura: una carica unitaria può essere creata dal nulla o scomparire nel vuoto. La carica elettrica complessiva di un sistema di particelle isolato non si può invece mai modificare e neppure il colore netto. La forza debole è inoltre caratterizzata dal suo raggio d'azione estremamente breve: i suoi effetti si estendono solo fino a distanze di circa 10^{-16} centimetri, pari all'incirca a un millesimo del diametro di un protone.

Nella teoria di gauge della forza debole sia la mancata conservazione della carica debole sia il breve raggio d'azione della forza sono attribuiti a un meccanismo chiamato rottura di simmetria spon-

La prima generazione di quark e di leptoni forma una configurazione ordinata se si dispongono le particelle secondo la loro carica elettrica. Sono rappresentati tutti i valori di carica da +1 a −1 in intervalli di 1/3. Tutte le particelle colorate hanno carica frazionaria, mentre quelle incolori hanno carica intera. La configurazione è un aspetto arbitrario del modello standard, in cui carica e colore sono indipendenti, ma potrebbe avere qualche spiegazione se i quark e i leptoni fossero composti.

tanea, che descriverò con maggiori particolari più oltre. Per ora è sufficiente sottolineare che il meccanismo di rottura della simmetria implica che la carica debole e la rotazione associata delle particelle debbano essere conservate a energia estremamente alta, dove la massa di una particella rappresenta una frazione trascurabile della sua energia cinetica.

La rottura spontanea della simmetria impone anche che i bosoni di gauge della forza debole siano particelle di grande massa; in realtà, essi hanno masse pari a circa 100 volte la massa del protone. Nel modello campione sono previsti tre di tali bosoni: due di essi, indicati con W^+ e W^-, posseggono sia carica elettrica sia carica debole; il terzo, chiamato Z^0, è elettricamente neutro. La grande massa dei bosoni deboli spiega il breve raggio d'azione della forza. Secondo il principio di indeterminazione della meccanica quantistica, il raggio d'azione di una forza è inversamente proporzionale alla massa della particella che la trasmette. Così, l'elettromagnetismo e la forza di colore, essendo trasportati da bosoni di gauge privi di massa, hanno in realtà un raggio d'azione infinito, mentre la forza debole ha una sfera d'influenza estremamente piccola. La rottura spontanea della simmetria ha un'ulteriore conseguenza: essa prevede l'esistenza di almeno un'altra particella di grande massa diversa dai bosoni deboli. Essa è chiamata particella di Higgs dal nome di Peter Higgs dell'Università di Edimburgo, che apportò un importante contributo alla teoria della rottura spontanea della simmetria.

Negli ultimi 10 anni i successi del modello standard hanno dato ai fisici una soddisfacente sicurezza. Tutte le forme conosciute di materia si possono costruire con i 18 quark colorati e con i sei leptoni del modello. Tutte le interazioni della materia osservate si possono spiegare come scambi dei 12 bosoni di gauge compresi nel modello: il fotone, gli otto gluoni e i tre bosoni deboli. Il modello pare coerente; nessuna parte di esso contrasta con altre e tutte le grandezze misurabili è previsto debbano avere un valore finito plausibile. La coerenza non è un risultato banale in un sistema concettuale di portata così ampia. Finora il modello risulta anche in accordo con tutti i risultati sperimentali, cioè a dire, nessuna chiara previsione del modello è stata ancora contraddetta dall'esperienza. A dire il vero, vi sono alcune importanti previsioni che non sono ancora state completamente verificate; in particolare, il neutrino tauonico, il quark alto, i bosoni deboli e la particella di Higgs devono ancora essere trovati. Sulla prima evidenza diretta dei bosoni W è stato riferito da un gruppo di ricercatori del CERN, l'Organizzazione europea di ricerche nucleari che ha sede a Ginevra (si veda in proposito *L'osservazione del bosone W* di Giorgio Salvini in «Le Scienze» n. 176, aprile 1983, p. 4). Negli anni a venire nuovi acceleratori di particelle e apparecchiature di rivelazione più sensibili verificheranno le restanti previsioni del modello. La maggior parte dei fisici sono del tutto sicuri che esse verranno confermate.

Se il modello standard si è dimostrato così valido, perché si dovrebbero prendere in considerazione teorie più elaborate? Il motivo principale non deriva dal sospetto che il modello standard sia sbagliato, ma piuttosto dalla sensazione che non sia completamente soddisfacente. Pur se il modello fornisce risposte corrette a tutte le domande che gli si rivolgono, molte domande sono rimaste senza risposta e molte regolarità della natura appaiono casuali o arbitrarie. In breve, il modello stesso richiede una spiegazione.

La prova più evidente dell'esistenza di qualche principio organizzatore che vada oltre il modello standard è fornita dalla proliferazione delle particelle elementari. Le proprietà conosciute della materia non sono così numerose o differenti da richiedere 24 particelle per rappresentarle tutte. In effetti, pare che vi siano molte ripetizioni nella gamma dei quark e dei leptoni. Vi sono tre leptoni con carica elettrica −1, tre leptoni neutri, tre quark con carica +2/3 e tre quark con carica −1/3. Ogni cosa viene triplicata senza alcun motivo apparente. Parrebbe che un mondo costruito scegliendo una sola particella da ciascuno dei quattro gruppi abbia tutta la necessaria varietà.

Da quanto risulta tutta la materia ordinaria può davvero essere formata da un sottoinsieme che comprenda soltanto il quark u, il quark d, l'elettrone e il neutrino elettronico. Queste quattro particelle con le loro antiparticelle costituiscono la «prima generazione» di quark e di leptoni. I restanti quark e leptoni ripetono semplicemente lo stesso schema in due ulteriori generazioni senza aggiungere in apparenza nulla di nuovo. Le particelle corrispondenti nelle diverse generazioni sono identiche sotto tutti gli aspetti tranne uno: esse hanno masse differenti. I quark d, s e b, per esempio, rispondono esattamente allo stesso modo alla forza elettromagnetica, a quella di colore e a quella debole. Tuttavia, per qualche motivo sconosciuto, il quark s è circa 20 volte più pesante del quark d e il quark b è circa 600 volte più pesante di quello d. Parimenti grandi e inspiegabili sono i rapporti tra le masse degli altri quark e dei leptoni carichi. (Le masse dei neutrini sono troppo piccole per essere state misurate; non si sa ancora se i neutrini sono solo molto leggeri o del tutto privi di massa.)

La presenza di tre generazioni di quark e di leptoni impone una spiegazione. Perché la natura si ripete? Anche la configurazione delle masse delle particelle è misteriosa. Nel modello standard le masse vengono determinate da circa 20 parametri «liberi» ai quali fu assegnato dai teorici qualsiasi valore a scelta; in pratica i valori si basano generalmente su scoperte sperimentali. È possibile che i 20 parametri siano tutti indipendenti? Sono costanti fondamentali della natura quali la velocità della luce o la carica elettrica dell'elettrone? Probabilmente no.

Un'altra regolarità sconcertante si osserva nelle cariche elettriche dei quark e dei leptoni: sono tutte correlate da rapporti semplici e sono tutte multipli interi di un terzo della carica dell'elettrone. Il modello standard non ne giustifica il motivo; in linea di principio i rapporti tra le cariche potrebbero avere qualsiasi valore. Dalle osservazioni si può ricavare che i rapporti un terzo e due terzi che definiscono le cariche dei quark non sono approssimazioni. Il protone è formato da due quark u e da un quark d, con cariche 2/3 + 2/3 − 1/3, cioè + 1. Se questi valori non fossero esatti e i quark avessero invece, per esempio, cariche +0,617 e −0,383, l'entità della carica del protone non coinciderebbe esattamente con quella dell'elettrone e gli atomi ordinari non sarebbero elettricamente neutri. Dal momento che gli atomi si possono raggruppare in quantità enormi, anche una lievissima deviazione dalla neutralità potrebbe essere subito rivelata.

Se si ordinano le particelle e le antiparticelle che formano una generazione secondo la loro carica, si trova che qualsiasi valore da −1 a +1 in intervalli di un terzo è occupato da una particella (o, nel caso di carica nulla, da due particelle, cioè il neutrino e l'antineutrino). Lo schema ottenuto pone altre domande. Perché la natura ha privilegiato questi valori di carica elettrica e non altri, quali +4/3 o −5/3? È evidente che tutte le particelle con carica intera sono incolori, mentre tutte quelle con carica frazionaria sono colorate. C'è qualche relazione tra la carica elettrica di una particella e il suo colore e tra i quark e i leptoni? Il modello standard non prevede tali relazioni, ma pare che esse esistano.

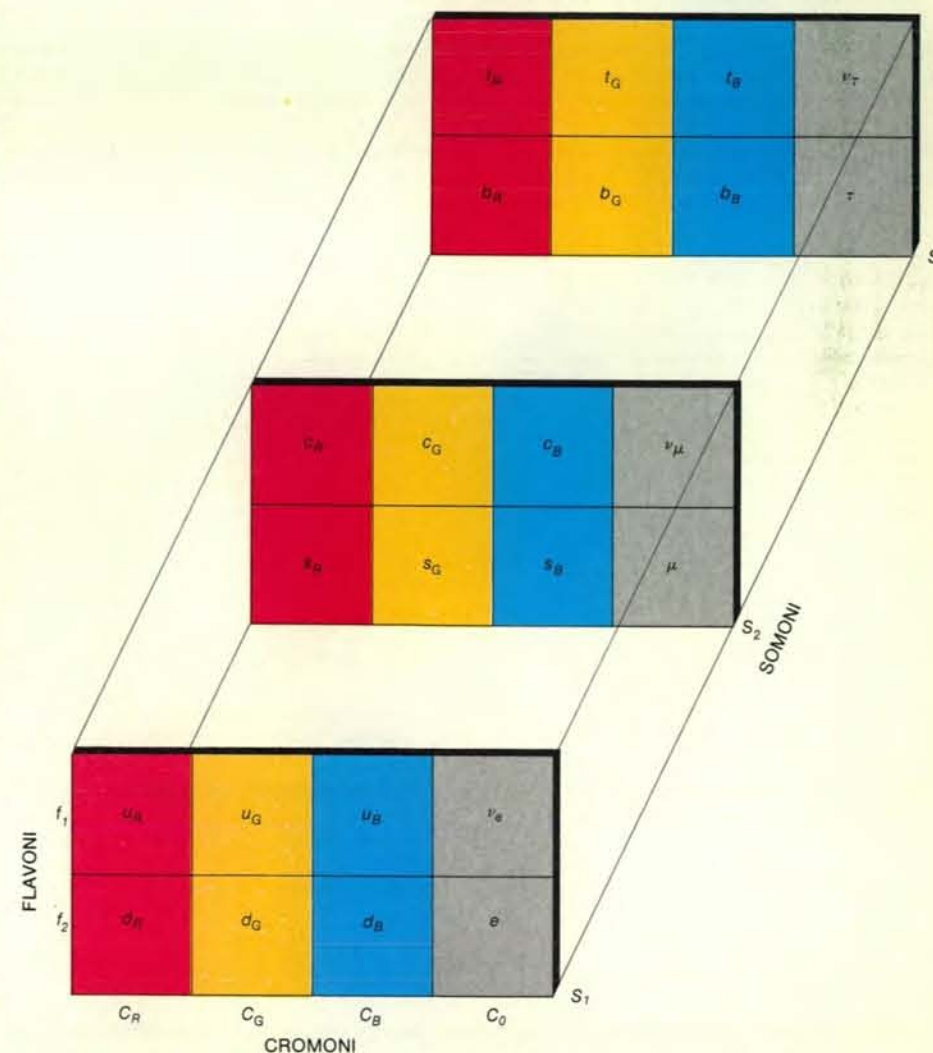
Un'altra motivazione per guardare al di là del modello standard è il desiderio continuo di unificare le forze fondamentali, o almeno di scoprire qualche relazione tra esse. Sarebbe per esempio servita la causa dell'economia se venissero unificate due di tali forze, come accadde per l'elettricità e il magnetismo, o se una forza potesse apparire come il residuo di un'altra, allo stesso modo in cui la forza forte diventò un residuo della forza di colore. Ironicamente, potrebbe risultare che una semplificazione di questo genere sia possibile soltanto introducendo altre forze.

Non è detto che una teoria che «va oltre» il modello standard debba contraddirlo o invalidarlo. Il modello standard può rivelarsi una buonissima approssimazione della teoria più approfondita. Il modello standard offre una descrizione molto efficace di tutti i fenomeni a distanze non inferiori a circa 10^{-16} centimetri. Una teoria più approfondita dovrebbe pertanto puntare su eventi di scala ancora più piccola. Se si devono scoprire nuovi costituenti, essi devono trovarsi all'interno di tali minuscole regioni di spazio. Se vi sono nuove forze, la loro azione deve essere efficace soltanto a distanza inferiore a 10^{-16} centimetri o perché la forza è per sua natura a breve raggio d'azione (seguendo l'esempio della forza debole) o perché è soggetta alla stessa forma di confinamento (come la forza di colore).

La ricerca di una teoria che vada oltre il modello standard fu suggerita quasi 10 anni fa e finora sono state esplorate diverse direzioni. Una direzione promettente ha portato ai modelli noti come teorie di grande unificazione, che comprendono in una sola forza fondamentale le forze elettromagnetica, di colore e debole. L'idea di base è quella di sistemare in una sola famiglia tutti i quark e i leptoni che formano una generazione; vengono quindi postulati nuovi bosoni di gauge per mediare le interazioni tra quark colorati e leptoni incolori. Le teorie spiegano le regolarità notate nella distribuzione della carica elettrica e spiegano la esatta commensurabilità delle cariche dei quark e dei leptoni. D'altra parte, esse non fanno nulla per ridurre il numero delle costanti fondamentali, non gettano luce sulla triplicazione delle generazioni e creano nuove difficoltà teoriche loro proprie.

Vi sono state molte variazioni sul tema della grande unificazione. Per esempio, il concetto di simmetria orizzontale supera il problema della triplicazione stabilendo una relazione di simmetria tra le generazioni. L'idea matematicamente elegante chiamata supersimmetria correla particelle il cui momento angolare di spin è semintero (quali i quark e i leptoni) a quelle con spin intero (quali i bosoni di gau-

	PREONE	CARICA ELETTRICA	COLORE	NUMERO DELLA GENERAZIONE
FLAVONI	f_1	+1/2	INCOLORE	0
	f_2	−1/2	INCOLORE	0
CROMONI	C_R	+1/6	ROSSO	0
	C_G	+1/6	GIALLO	0
	C_B	+1/6	BLU	0
	C_0	−1/2	INCOLORE	0
SOMONI	S_1	0	INCOLORE	1
	S_2	0	INCOLORE	2
	S_3	0	INCOLORE	3

Il modello a preoni assegna tre proprietà dei quark e dei leptoni a tre gruppi di ipotetici costituenti chiamati flavoni, cromoni e somoni. Un quark o un leptone si formano prelevando un preone da ciascun gruppo. I flavoni sono i principali responsabili della determinazione della carica elettrica, i cromoni determinano il colore e i somoni determinano il numero della generazione. Idealmente ciascun tipo di preone dovrebbe trasportare soltanto una proprietà, ma si rende necessario qualche aggiustamento per differenziare le cariche elettriche frazionarie dei quark da quelle intere dei leptoni. Nella versione del modello qui descritta i cromoni trasportano carica elettrica e colore.

Combinazioni di preoni danno origine ai 24 quark e leptoni delle tre generazioni. Per esempio, il quark rosso s è formato dal somone S_2 (notazione che indica che il composto è una particella della seconda generazione), in combinazione con il cromone rosso e con il flavone negativo.

ge). La teoria del technicolor suggerisce che la particella di Higgs del modello standard sia un oggetto composto, costituito di nuove entità fondamentali; esse dovrebbero essere tenute assieme da una nuova forza analoga alla forza di colore e chiamata technicolor. Ciascuna di queste idee dà una risposta ad alcuni problemi ancora aperti del modello standard. Ciascuna idea non è però capace di rispondere ad altre domande, pone nuovi problemi e peggiora quelli esistenti, aumentando per esempio ulteriormente il numero di costanti arbitrarie non correlate.

In tutti gli schemi per la grande unificazione sopra indicati, si assume esplicitamente che i quark, i leptoni, il fotone, i gluoni e i bosoni deboli siano le particelle veramente fondamentali della teoria definitiva della natura. L'alternativa di suggerire che i quark e i leptoni siano a loro volta composti è in un certo senso l'ipotesi più conservatrice e meno originale. Si tratta di una strategia che ha funzionato ripetutamente in passato nell'andare dall'atomo al nucleo al protone al quark. In un altro senso l'idea della sottostruttura dei quark e dei leptoni è una proposta molto radicale. L'elettrone si studia oggi da quasi un secolo e la sua natura puntiforme è stata accertata molto chiaramente. Nel caso del neutrino, che potrebbe risultare del tutto privo di massa, è ancora più difficile pensare a una struttura interna. L'affermazione che queste particelle e le altre a esse simili siano composte dovrà ovviamente superare ostacoli formidabili per avere un futuro.

Trascurando le difficoltà dell'impresa, rimangono le sue potenzialità. Un modello composto di completo successo potrebbe risolvere tutti i problemi lasciati insoluti dal modello standard. Una teoria ipotetica di questo genere dovrebbe cominciare introducendo un nuovo gruppo di particelle elementari, che chiamerò genericamente prequark. In teoria non dovrebbero essercene molti. Tutti i quark e i leptoni del modello standard potrebbero essere spiegati come una combinazione di prequark, allo stesso modo in cui ciascun adrone si può spiegare come una combinazione di quark. La massa di un quark e di un leptone non sarebbe più una costante arbitraria della natura, ma sarebbe determinata dalla massa dei prequark costituenti e dall'intensità della forza che tiene assieme i prequark. Allo stesso modo si potrebbero spiegare i rapporti esatti che correlano la carica di un quark a quella di un leptone: entrambi i tipi di particelle composte ricaverebbero la loro carica da quella degli stessi prequark costituenti. La struttura completa dei quark e dei leptoni di una generazione rifletterebbe forse alcune regole semplici per combinare i prequark.

Anche l'esistenza di più generazioni si potrebbe spiegare in modo naturale. I quark e i leptoni delle generazioni di livello superiore potrebbero presentare una costituzione interna simile a quella delle corrispondenti particelle della prima generazione; le differenze potrebbero riscontrarsi nell'energia e nello stato di moto dei costituenti. Così i quark s e b sarebbero stati eccitati del quark d e il muone e il leptone tau sarebbero stati eccitati dell'elettrone. Stati eccitati del genere si conoscono per tutti gli altri sistemi composti, tra i quali gli atomi, i nuclei e gli adroni. Per esempio, negli esperimenti almeno una decina di adroni sono stati identificati come stati eccitati del protone; si pensa che questi e il protone stesso abbiano essenzialmente la stessa composizione di quark, cioè *uud*.

Il modello a rishoni ricava tutti i quark e i leptoni partendo da due soli tipi di particelle fondamentali e dalle corrispondenti antiparticelle. I rishoni trasportano sia l'ipercolore, una proprietà associata alla forza che li lega l'uno all'altro, sia il colore ordinario, che essi trasferiscono ai sistemi composti che formano. Un rishone è elettricamente carico, mentre l'altro è neutro.

COMBINAZIONE DI RISHONI	PARTICELLA	COLORE	CARICA ELETTRICA
TTT	$\bar{e}$	INCOLORE	+1
TTV	u	ROSSO GIALLO BLU	+2/3
TVV	$\bar{d}$	ANTIROSSO ANTIGIALLO ANTIBLU	+1/3
VVV	ν_e	INCOLORE	0
$\overline{VVV}$	$\bar{\nu}_e$	INCOLORE	0
$\overline{VVT}$	d	ROSSO GIALLO BLU	−1/3
$\overline{VTT}$	$\bar{u}$	ANTIROSSO ANTIGIALLO ANTIBLU	−2/3
$\overline{TTT}$	e	INCOLORE	−1

Combinazioni di rishoni a gruppi di tre forniscono una corretta spiegazione di tutti i quark e i leptoni (e dei corrispondenti antiquark e antileptoni) di qualsiasi generazione. La configurazione delle cariche elettriche osservata nel modello standard e la evidente relazione tra carica frazionaria e colore appaiono come naturali conseguenze del modo in cui si combinano i rishoni. Tutte le combinazioni permesse di tre rishoni o di tre antirishoni sono neutre rispetto all'ipercolore.

Questa teoria immaginaria ideale dei prequark raggiunge tutti gli obiettivi voluti tranne l'unificazione delle forze fondamentali. Perfino a questo proposito si spera di compiere qualche passo avanti, dato che dovrebbe essere molto probabile l'introduzione di una nuova forza che tenga assieme i prequark; la nuova forza potrebbe portare a nuove conoscenze sulla correlazione delle forze note. Tuttavia, immaginare a cosa potrebbe rassomigliare un modello di successo non è affatto lo stesso che costruirne davvero uno realistico e coerente. Finora nessuno lo ha fatto.

Ciò che manca è una teoria soddisfacente della dinamica dei prequark, una teoria che descriva come si muovono i prequark all'interno di un quark o di un leptone e che permetta di calcolare la massa e l'energia totale del sistema. Come preciserò in seguito, esistono ostacoli fondamentali alla formulazione di una tale teoria, anche se a mio parere non sarebbero insuperabili. Nel frattempo, in mancanza di qualsiasi spiegazione convincente del moto dei prequark, i teorici stanno ugualmente esplorando le possibilità combinatorie dell'idea dei prequark, stanno cioè esaminando i modi in cui si potrebbero costruire quark e leptoni come combinazioni specifiche di costituenti più fini.

Negli ultimi anni sono state proposte diverse decine di modelli composti; essi si possono classificare forse in quattro o cinque gruppi principali. Nessuno dei modelli risolve tutti i problemi, risponde a tutte le domande ed è universalmente accettato. Sarebbe scorretto descrivere soltanto uno schema, ma è impossibile enumerarli tutti. Presenterò alcune delle idee fondamentali.

Il primo modello esplicito di una sottostruttura dei quark e dei leptoni è stato proposto nel 1974 da Jogesh C. Pati dell'Università del Maryland a College Park e da Salam, i quali da allora sono ritornati più volte sull'argomento in collaborazione con John Strathdee del Centro internazionale di fisica teorica. Sono stati loro a introdurre il termine prequark, che io ho adottato in questa sede come nome generico per ipotetici subcostituenti di qualsiasi genere. Chiamerò «preoni» le particelle elementari specifiche del modello previste da Pati e Salam, altro termine di loro invenzione.

La razionalità del modello a preoni comincia con l'osservazione che tutti i quark e i leptoni si possono identificare in modo inequivocabile elencando soltanto tre delle loro proprietà: la carica elettrica, il colore e il numero della generazione. Tali proprietà suggeriscono poi un modo immediato di costruire un gruppo di particelle costituenti. Sono necessarie tre famiglie di preoni. In una famiglia i preoni hanno carica elettrica, in un'altra trasportano il colore e nella terza hanno qualche proprietà che determina il numero della generazione. Qualsiasi quark o leptone si costruisce scegliendo esattamente un preone da ogni famiglia.

I preoni che determinano il numero della generazione sono detti «somoni», dal greco *soma*, che significa corpo, perché essi hanno un effetto predominante sulla massa del sistema composto. Dal momento che esistono tre generazioni di quark e di leptoni, devono esistere tre somoni. Il colore del sistema composto viene determinato da preoni chiamati «cromoni»; ne esistono quattro, uno rosso, uno giallo, uno blu e uno incolore. La restante famiglia di preoni, a cui è assegnato il ruolo di definire la carica elettrica, richiede solo due membri, in modo da poter identificare in modo univoco tutti i quark e i leptoni. A questi ultimi preoni è stato dato il nome di «flavoni», da *flavor* (sapore), il misterioso termine che indica qualsiasi proprietà che contraddistingue il quark u dal quark d, il c dall's, il neutrino dall'elettrone e così via.

Nel modello a preoni la classificazione di una particella composta deriva direttamente dall'insieme dei suoi preoni. Per esempio, tutti i leptoni sono caratterizzati da un cromone incolore e tutte le particelle della prima generazione devono ovviamente avere un somone della prima generazione. Nell'attribuzione della carica elettrica nasce invece una complicazione. Se esistono soltanto due flavoni e se essi sono i soli portatori di carica elettrica, non si possono riprodurre tutti i valori di carica osservati in natura. Per esempio, il quark u e il neutrino devono avere la stessa carica (perché comprendono lo stesso flavone) e per la stessa ragione il quark d e l'elettrone. La difficoltà si può superare in molti modi. In uno schema viene assegnata una carica elettrica sia ai flavoni sia ai cromoni e la carica totale di una particella composta risulta uguale alla somma dei due valori. Si possono costruire modelli del genere per fornire il valore corretto degli stati di carica, ma solo a patto di abbandonare il principio che ogni tipo di preone trasporti solo una proprietà.

Un altro preoccupante aspetto del modello a preoni è il requisito che le particelle composte si possano formare solo estraendo un preone per famiglia. Perché non esistono, per esempio, particelle formate da tre cromoni o da due somoni e un flavone? Le proprietà particolari di tali particelle le renderebbero estremamente numerose. Pare probabile che, se fossero esistite, sarebbero già state rivelate.

Altri fisici hanno proposto molte varianti del modello a preoni, utilizzando la stessa idea fondamentale, ma gruppi di preoni lievemente diversi. Tra le varianti sono degni di nota i modelli suggeriti da Hidezumi Terazawa, Yoichi Chikashige e Keiichi Akama dell'Università di Tokio e da O. Wallace Greenberg e Joseph Sucher dell'Università del Maryland.

Il modello forse più semplice della struttura dei quark e dei leptoni è il modello a rishoni, da me proposto nel 1979. Un'idea del genere è stata avanzata più o meno nello stesso periodo da Michael A. Shupe dell'Università dell'Illinois a Urbana-Champaign. Il modello è stato da allora ulteriormente perfezionato e studiato in grande dettaglio da Nathan Seiberg e da me al Weizmann Institute of Science a Rehovot. Il modello postula soltanto due specie di mattoni fondamentali chiamati rishoni. (*Rishon* è l'aggettivo ebraico che sta per primario o fondamentale.) Un rishone ha carica elettrica + 1/3 e l'altro è elettricamente neutro. Li ho chiamati rispettivamente T e V da *Tohu Vavohu*, espressione ebraica per «informe e vuoto», la descrizione dello stato iniziale dell'universo data nel primo capitolo della *Genesi*. Gli antirishoni complementari hanno cariche −1/3 e zero e sono indicati con $\bar{T}$ e $\bar{V}$.

Il modello per costruire un quark o un leptone segue una regola semplice: tre rishoni qualsiasi, o tre qualsiasi antirishoni, possono essere raggruppati per formare un sistema composto, ma in una sola particella non possono essere mescolati rishoni e antirishoni. La regola dà origine a 16 combinazioni, che riproducono esattamente le proprietà dei 16 quark, antiquark, leptoni e antileptoni della prima generazione. In altri termini, ogni quark e leptone della prima generazione corrisponde a certe combinazioni permesse di rishoni o di antirishoni. (In questo sistema di classificazione ogni colore viene contato separatamente.)

La configurazione delle cariche dei quark e dei leptoni viene costruita nel modo seguente. La combinazione TTT, con cariche rishoniche di 1/3 + 1/3 + 1/3, ha una carica totale + 1 e corrisponde quindi al positone; allo stesso modo, $\overline{TTT}$ ha una carica totale − 1 e viene identificato con l'elettrone. Le combinazioni VVV e $\overline{VVV}$ sono entrambe elettricamente neutre e rappresentano rispettivamente il neutrino e l'antineutrino. Le altre combinazioni permesse portano a quark dotati di carica frazionaria. TTV, con una carica + 2/3, è il quark u, mentre TVV, con una carica + 1/3, è l'antiquark $\bar{d}$. Gli analoghi stati di antirishone, $\overline{VVT}$ e $\overline{VTT}$, corrispondono al quark d e all'antiquark $\bar{u}$.

Il modello spiega anche con successo il colore dei sistemi composti. Un rishone T può avere uno qualsiasi dei tre colori rosso, giallo e blu, mentre un rishone V ha un anticolore. Combinazioni quali $\overline{TTT}$ e VVV, che designano leptoni, si possono rendere incolori, dal momento che comprendono un rishone di ciascun colore o uno di ciascun anticolore. Le altre combinazioni, che danno origine a quark, devono avere un colore risultante. Per esempio, uno stato TTV potrebbe avere colori di rishone rosso, blu e antiblu; l'antiblu cancellerebbe il blu, lasciando il sistema con un colore risultante rosso. In questo modo viene chiaramente evidenziata la connessione tra colore e carica elettrica, intuita, ma non spiegata dal modello standard. Per il modo in cui ai rishoni vengono assegnati carica elettrica e colore, tutti i sistemi composti con carica frazionaria risultano colorati, mentre tutti i sistemi con carica elettrica intera si possono rendere incolori.

Anche altre regolarità del modello standard perdono la loro aria di mistero allorché si introducono i rishoni. Si consideri l'atomo di idrogeno, formato da un

protone e da un elettrone, o, in termini di quark e di leptoni, da due quark *u*, un quark *d* e un elettrone. Il contenuto totale di rishoni nei quark è di quattro *T*, un $\bar{T}$, due *V* e due $\bar{V}$. La carica elettrica di $\bar{T}$ annulla la carica di un rishone *T*, e anche *V* e $\bar{V}$ si annullano (essi sono comunque privi di carica), lasciando il protone con una carica complessiva uguale a quella di un sistema *TTT*. Il contenuto in rishoni dell'elettrone è l'opposto: $\overline{TTT}$. È quindi evidente perché il protone e l'elettrone abbiano cariche di ugual modulo e perché l'atomo di idrogeno sia neutro: le fonti ultime della carica sono coppie di particelle e di antiparticelle accoppiate.

Il modello a rishoni e molti altri modelli che spiegano la prima generazione incontrano difficoltà nello spiegare la seconda e la terza generazione. Sembrerebbe che tali modelli siano adatti essi stessi allo schema di formazione di ogni particella delle generazioni di livello superiore come stato eccitato della corrispondente particella della prima generazione. L'idea più semplice sarebbe la descrizione del muone fatta, per esempio, con gli stessi prequark che formano l'elettrone, ma con una configurazione di energia più elevata per i prequark del muone. Si tratta di un'idea elegante ma, purtroppo, poco pratica. Lo schema implica differenze di energia tra i successivi stati eccitati molto maggiori delle differenze reali. Il difetto è fondamentale e pare non vi sia rimedio.

Sono stati presi in esame altri possibili meccanismi di creazione di generazioni multiple. Numerosi fisici hanno suggerito che le particelle di un dato stato nelle generazioni di livello superiore si potrebbero creare aggiungendo una particella di Higgs, la particella «extra» associata ai bosoni deboli nel modello standard. Dato che una particella di Higgs non possiede né carica elettrica né colore e neppure momento angolare di spin, l'aggiunta di una di esse a un sistema composto ne modificherebbe soltanto la massa. Quindi un elettrone si potrebbe trasformare in un muone aggiungendogli una particella di Higgs o in una tau aggiungendogli due o più particelle di Higgs. Seiberg e io abbiamo proposto un altro possibile meccanismo: si potrebbe formare una particella di una generazione di livello maggiore mediante l'aggiunta di coppie di prequark e di antiprequark. In una coppia siffatta tutte le cariche e le altre proprietà dovrebbero annullarsi e in tal modo ne sarebbe modificata ancora una volta soltanto la massa.

Queste idee sono attualmente in fase di libera speculazione. Nessuno sa che cosa distingua le tre generazioni l'una dall'altra o perché ve ne siano tre o se ce ne possano essere di più. Non si può dare alcuna spiegazione della differenza di massa tra le generazioni. In breve, la triplicazione delle generazioni è tuttora uno dei più grandi enigmi irrisolti.

Merita di essere citato un terzo tipo di modello di sottostruttura. Esso cerca di collegare la possibilità di una struttura di quark e di leptoni a un altro problema fondamentale: la comprensione della teoria quantistica relativistica della gravitazione. Idee di questo genere sono state esplorate da John Ellis, Mary K. Gaillard, Luciano Maiani e Bruno Zumino del CERN. Un modo di affrontare le loro idee è quello di considerare le distanze alle quali interagiscono i prequark: il limite sperimentale è inferiore a 10^{-16} centimetri, ma la distanza reale potrebbe essere di diversi ordini di grandezza inferiore. A circa 10^{-34} centimetri la forza gravitazionale diventa sufficientemente intensa da avere un effetto significativo sulle singole particelle. Se la scala delle interazioni tra i prequark è così piccola, non si può trascurare la gravitazione. Ellis, Gaillard, Maiani e Zumino hanno tracciato un ambizioso progetto che si propone di unificare tutte le forze, compresa la gravitazione, in uno schema che tratta come particelle composte non soltanto i quark e i leptoni, ma anche i bosoni di gauge. Anche questo modello, come altri modelli composti, presenta serie pecche.

Qualsiasi modello a prequark, indipendentemente dai suoi particolari, deve fornire un qualche meccanismo per tener legati i prequark, tra i quali deve esistere una potente forza attrattiva. Una strategia è quella di postulare una nuova forza fondamentale della natura analoga nelle sue azioni alla forza di colore del modello standard. Per evidenziare l'analogia la nuova forza è detta «forza di ipercolore» e i campi portatori sono detti «ipergluoni». Si ipotizza che i prequark abbiano ipercolore, ma che essi si combinino formando sistemi composti senza ipercolore allo stesso modo in cui i quark, dotati di colore ordinario, si combinano e formano protoni e neutroni incolori. La forza di ipercolore probabilmente dà anche origine alla proprietà del confinamento, ancora in analogia alla forza di colore. Quindi tutti i prequark ipercolorati sarebbero intrappolati all'interno di particelle composte, il che spiegherebbe perché non si osservano prequark liberi negli esperimenti. Un'idea del genere è stata proposta per primo da 't Hooft che studiò alcune delle sue conseguenze matematiche, ma che espresse anche dubbi sul fatto che la natura si comporti davvero così.

Il tipico raggio di confinamento dell'ipercolore deve essere inferiore a 10^{-16} centimetri. Solo sondando la materia a distanze inferiori a quest'ultima sarebbe possibile osservare gli ipotetici prèquark e i loro ipercolori. A una distanza di 10^{-14} o 10^{-15} centimetri l'ipercolore scompare realmente; i soli oggetti osservabili a tale scala di risoluzione (i quark e i leptoni) sono neutri rispetto all'ipercolore. A una distanza di 10^{-13} centimetri anche il colore ordinario svanisce e il mondo appare formato interamente da oggetti privi sia del colore sia dell'ipercolore: protoni, neutroni, elettroni e così via.

Le dimensioni e l'energia sono correlate nella teoria dei quanti, il che indica che i costituenti di quark e di leptoni composti devono avere un'energia cinetica eccezionalmente elevata. Le dimensioni di un atomo implicano che i suoi costituenti possano avere energie che variano nell'intervallo compreso tra qualche elettronvolt e qualche migliaio di elettronvolt. (Un elettronvolt è l'energia acquistata da un elettrone accelerato dalla differenza di potenziale di un volt.) In un nucleo i protoni e i neutroni si muovono con un'energia di parecchi milioni di elettronvolt e in un protone o in un neutrone i quark hanno energie di parecchie centinaia di milioni di elettronvolt. Qualsiasi costituente dei quark e dei leptoni deve essere confinato in un raggio inferiore a 10^{-16} centimetri e possibilmente molto inferiore. L'energia cinetica degli ipotetici prequark non potrebbe dunque essere inferiore a qualche centinaio di miliardi di elettronvolt.

La nozione di ipercolore si adatta bene a una varietà di modelli a prequark, tra i quali il modello a rishoni. Si suppone che, oltre alla carica elettrica e al colore, i rishoni abbiano ipercolore e gli antirishoni antiipercolore. Sono permesse soltanto combinazioni di tre rishoni o di tre antirishoni perché soltanto combinazioni del genere sono neutre rispetto all'ipercolore. Un sistema misto di tre particelle del tipo $TT\bar{T}$ non può esistere perché esso non sarebbe senza ipercolore. L'assegnazione degli ipercolori spiega perciò la regola per formare sistemi composti da rishoni. Analoghe regole valgono per altri modelli a prequark basati sull'ipercolore.

Se l'obiettivo di un modello a prequark è quello di semplificare la comprensione della natura, il postulare una nuova forza fondamentale non pare sia di molto aiuto. Nel caso dell'ipercolore, però, potrebbe esservi qualche vantaggio. Si consideri il neutrino: esso non ha né carica elettrica né colore, ma solo carica debole. Secondo il modello standard, due neutrini possono interagire soltanto attraverso la forza debole a breve raggio d'azione. Se i neutrini sono composti da prequark ipercolorati, però, potrebbe esistere una sorgente addizionale di interazioni tra i neutrini. Quando due neutrini sono molto lontani, in pratica non vi sono tra essi forze di ipercolore, ma quando essi sono a breve distanza, i prequark ipercolorati all'interno di un neutrino sono in grado di «vedere» gli ipercolori interni dell'altro. Il risultato è una serie di complesse attrazioni e repulsioni a breve raggio d'azione. Il meccanismo è ovviamente esattamente lo stesso di quello che spiega la forza molecolare come residuo della forza elettromagnetica e la forza forte come residuo della forza di colore.

Anche la conclusione può essere la stessa. Seiberg e io e, indipendentemente, Greenberg e Sucher, fummo i primi a suggerire che la forza debole a breve raggio d'azione potesse essere in realtà un effetto residuo della forza di ipercolore. Secondo questa ipotesi, anche i bosoni deboli W^+, W^- e Z^0 devono essere oggetti composti forse costituiti da certe combinazioni degli stessi prequark che compongono i quark e i leptoni. Se tale idea verrà confermata, l'elenco delle forze fondamentali avrà ancora quattro lemmi: gravitazione, elettromagnetismo, colore e ipercolore. Si deve però far notare che tutte queste forze sono a lungo raggio d'azione; le forze a breve raggio d'azione molecolare, forte e debole avrebbero perduto la loro caratteristica fondamentale.

Il paradosso della massa e dell'energia rende difficile arrivare a una teoria sul moto e sulle interazioni dei prequark. In un atomo o in un nucleo l'energia cinetica dei costituenti *(in colore)* è molto inferiore alla massa totale del sistema *(in grigio)*. In un protone le due grandezze sono di valore paragonabile. In un quark composto invece l'energia dei prequark supera di molto la massa totale. In realtà, se confrontata con l'energia cinetica, la massa è praticamente nulla. In un modo o nell'altro la massa si annulla praticamente tutta, evento che difficilmente è accidentale.

Per ora l'ipercolore rimane un'ipotesi e altrettanto vale per l'idea di spiegare la forza debole come un residuo della forza di ipercolore. Può anche risultare che la forza debole è fondamentale. Una accurata misura della massa, della vita media e di altre proprietà dei bosoni deboli dovrebbe chiarire il problema.

L'ipercolore non è il solo candidato per una forza di legame tra i prequark. Un'altra interessante possibilità è stata suggerita da Pati, Salam e Strathdee. Anziché introdurre una nuova forza di ipercolore, essi hanno preso in prestito un'idea da tempo familiare, cioè la forza magnetica, e l'hanno adattata a un nuovo scopo. Un magnete ordinario ha invariabilmente due poli, che si possono pensare di cariche magnetiche opposte. Da 50 anni esistono motivi teorici per supporre che possano esistere anche cariche magnetiche isolate, o monopoli. Pati, Salam e Strathdee hanno dedotto che i prequark potrebbero essere particelle con cariche somiglianti sia a cariche magnetiche sia a cariche elettriche. Se è così, le forze che li legano potrebbero avere una nuova e interessante origine.

Nessuna delle idee che ho appena descritto costituisce una teoria della dinamica dei prequark. Esiste infatti un serio ostacolo alla formulazione di una tale teoria; è il requisito che i prequark siano estremamente piccoli. Il limite più rigoroso alle loro dimensioni è fornito indirettamente dalle misurazioni del momento magnetico dell'elettrone, che è in accordo con i calcoli dell'elettrodinamica quantistica con una precisione di 10 cifre significative. Nei calcoli si suppone che l'elettrone sia puntiforme; se avesse una qualsiasi estensione spaziale o una struttura interna, il valore misurato differirebbe da quello calcolato. Evidentemente una qualsiasi discrepanza del genere può al massimo influenzare la undicesima cifra del risultato. È questo vincolo a limitare la scala di distanze caratteristica della struttura interna dell'elettrone al di sotto di 10^{-16} centimetri. Per parlar chiaro, questo è il massimo raggio di un elettrone e qualsiasi prequark deve trovarsi all'interno di esso. Se essi fossero sparpagliati su un raggio maggiore, la loro presenza sarebbe già stata rivelata.

Perché le piccole dimensioni dell'elettrone dovrebbero vietare illazioni sulla

sua struttura interna? Il principio di indeterminazione stabilisce una relazione reciproca tra le dimensioni di un sistema composto e l'energia cinetica di qualsiasi componente in moto al suo interno. Più piccolo è il sistema, maggiore è l'energia cinetica dei costituenti. Ne deriva che i prequark devono avere energie enormi: più di 100 GeV (100 miliardi di elettronvolt), e forse molto di più. (Un elettronvolt è l'energia acquistata da un elettrone accelerato dalla differenza di potenziale di un volt.) Dal momento che la massa è fondamentalmente equivalente all'energia, si può misurare nello stesso sistema di unità. Per esempio, la massa di un elettrone è equivalente a 0,0005 GeV. Qui nasce un paradosso, che chiamerò paradosso energetico: la massa del sistema composto (se è davvero composto) è molto minore dell'energia dei suoi costituenti.

Si può chiarire la stranezza della situazione considerando le relazioni tra la massa e l'energia cinetica di altri sistemi composti. In un atomo l'energia cinetica di un elettrone tipico è inferiore alla massa dell'atomo per molti ordini di grandezza. Per esempio, nell'idrogeno il rapporto è di circa una parte su 100 milioni. L'energia necessaria per cambiare l'orbita dell'elettrone e per portare perciò l'atomo in uno stato eccitato è allo stesso modo una frazione trascurabile della massa atomica. In un nucleo l'energia cinetica dei protoni e dei neutroni è anch'essa piccola in confronto alla massa del nucleo, ma non è completamente trascurabile. Il moto delle particelle conferisce loro un'energia equivalente a circa l'1 per cento della massa del sistema. Anche l'energia necessaria per creare uno stato eccitato è circa l'1 per cento della massa.

Per il protone e i suoi quark costituenti la relazione massa-energia comincia a diventare curiosa. Noto il raggio efficace del protone, si può calcolare l'energia tipica dei suoi quark costituenti; essa risulta paragonabile alla massa dello stesso protone, che è di poco inferiore a 1 GeV. L'energia che si deve impiegare per creare uno stato eccitato del sistema dei quark è dello stesso ordine di grandezza: gli adroni identificati come stati eccitati del protone hanno una massa superiore a quella del protone di un valore compreso tra il 30 e il 100 per cento. Ciononostante, il rapporto tra l'energia cinetica e la massa totale è ancora nell'intervallo che pare intuitivamente ragionevole. Supponiamo di conoscere solo il raggio del protone, e quindi la tipica energia di tutto ciò che può trovarsi al suo interno, e che ci venga chiesto di indovinare la massa del protone. Essendo l'energia dei costituenti generalmente di poche centinaia di milioni di elettronvolt, risponderemmo certamente che la massa totale del sistema è almeno dello stesso ordine di grandezza e forse maggiore. La risposta sarebbe corretta.

Per l'atomo, il nucleo e il protone, allora, la massa del sistema è almeno altrettanto grande dell'energia cinetica dei costituenti e in alcuni casi è molto maggiore. Tuttavia, se i quark e i leptoni sono composti, la relazione tra l'energia e la massa deve essere del tutto diversa. Dato che i prequark hanno energie ben superiori a 100 GeV, si concluderebbe che essi formano sistemi composti con masse di centinaia di GeV o oltre. In realtà i quark e i leptoni noti hanno masse molto minori; nel caso dell'elettrone e dei neutrini la massa è inferiore di almeno sei ordini di grandezza. Il tutto è molto minore della somma delle sue parti.

Anche l'elevata energia dei prequark è un ostacolo all'idea di considerare le generazioni di livello superiore dei quark e dei leptoni come stati eccitati dello stesso insieme di prequark che formano le particelle della prima generazione. Come negli altri sistemi composti, l'energia necessaria per cambiare le orbite dei prequark dovrebbe essere dello stesso ordine di grandezza dell'energia cinetica dei costituenti. Si dovrebbe perciò prevedere che le generazioni successive differiscano in massa di centinaia di GeV, mentre le differenze di massa effettive sono solo di circa 0,1 GeV.

A questo punto si potrebbe benissimo adottare l'interpretazione che il paradosso dell'energia non possa essere accettato, dal momento che esso dimostra semplicemente la natura elementare e priva di struttura dei quark e dei leptoni. Molti fisici seguono questa interpretazione. Tuttavia, il paradosso dell'energia non contraddice alcun principio fondamentale della fisica, e potrei concludere che le prove circostanziali a favore di una struttura composta dei quark e dei leptoni sono sufficientemente persuasive da incoraggiare ulteriori ricerche.

Ciò che è peculiare delle masse dei quark e dei leptoni non è semplicemente il fatto che siano piccole, ma che esse siano virtualmente nulle se misurate nella scala energetica definita dall'energia dei loro costituenti. In altri sistemi composti una piccola quantità di massa viene «perduta» per conversione in energia di legame del sistema. Per esempio, la massa totale di un atomo di idrogeno è leggermente inferiore a quella di un protone e di un elettrone isolati; la differenza è uguale all'energia di legame. In un nucleo questo «difetto di massa» può raggiungere qualche per cento della massa totale. Sembra che in un quark o in un leptone l'intera massa del sistema si annulli quasi completamente. Un annullamento «miracoloso» del genere non è sicuramente impossibile, ma pare molto improbabile che sia casuale. Grandi annullamenti di questo tipo sono noti altrove in fisica, ed è sempre risultato che derivano da qualche principio di simmetria o di conservazione. Per aver la speranza di costruire una teoria della dinamica dei prequark, è essenziale scoprire una simmetria del genere anche in questo caso.

Esiste un probabile candidato, la simmetria chirale, o chiralità. Il nome deriva dal vocabolo greco che sta per mano, e la simmetria ha a che fare con il carattere destrorso o sinistrorso, la proprietà definita dallo spin e dalla direzione di moto di una particella. Come altre simmetrie della natura, la simmetria chirale gode di un principio di conservazione che spiega con la massima chiarezza ciò che la simmetria significa. Tale principio afferma che la quantità totale di particelle destrorse e la quantità totale di particelle sinistrorse non possono mai variare.

Nel mondo comune dei protoni, degli elettroni e di particelle simili il carattere destrorso o sinistrorso o chiralità manifestamente non si conserva. Una violazione del principio di conservazione si può dimostrare con un semplice esperimento ideale. Si supponga che un osservatore si muova di moto rettilineo quando viene superato da un elettrone. Quando l'elettrone si allontana, l'osservatore nota che il suo spin e la sua direzione di moto sono correlati dalla regola della mano destra. Supponiamo ora che l'osservatore acceleri, in modo da superare l'elettrone. Nel sistema di riferimento dell'osservatore pare che l'elettrone si stia avvicinando; in altri termini, ha invertito la direzione. Tuttavia, dal momento che il suo spin non è cambiato, è diventato una particella sinistrorsa.

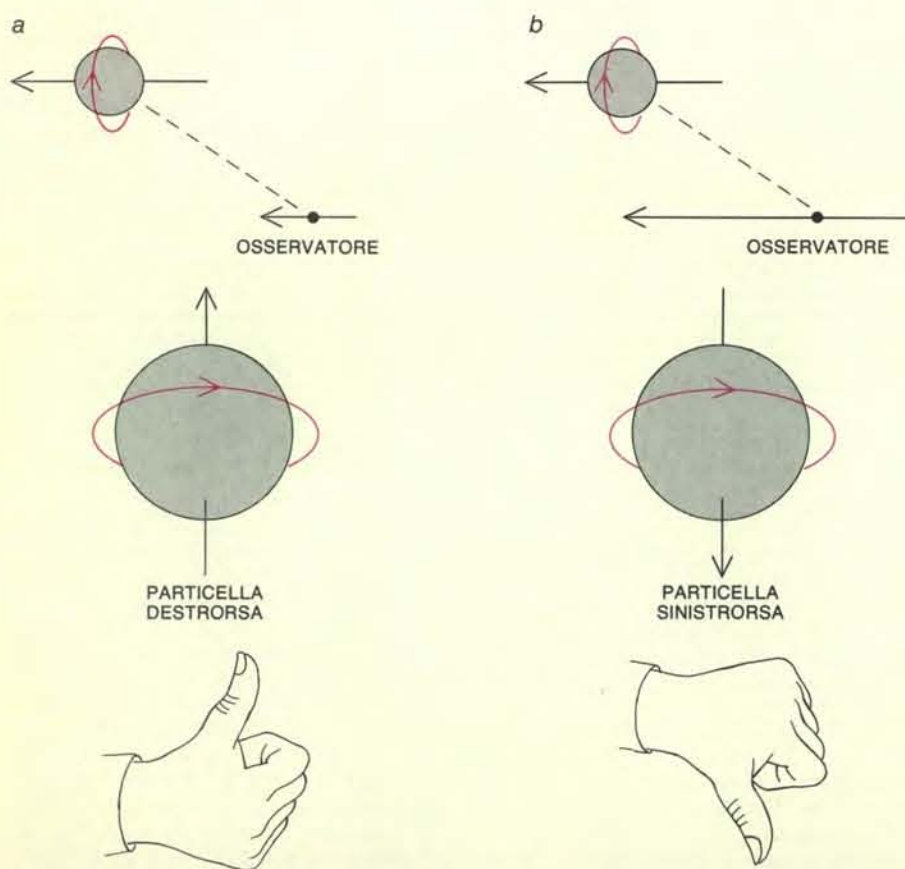
Vi è un tipo di particella a cui questo esperimento ideale non può essere applicato: una particella priva di massa. Dato che una particella priva di massa deve sempre muoversi alla velocità della luce, nessun osservatore può mai superarla in velocità. Il carattere destrorso o sinistrorso di una particella priva di massa è dunque una proprietà invariante, indipendentemente dal sistema di riferimento dell'osservatore. Inoltre, si può dimostrare che nessuna delle forze della natura conosciute (quelle mediate dai fotoni, dai gluoni e dai bosoni deboli) può modificare il carattere destrorso o sinistrorso di una particella. Così, se il mondo fosse costituito esclusivamente da particelle prive di massa, si potrebbe affermare che esso ha simmetria chirale.

La simmetria chirale è alla base di un'idea che potrebbe spiegare ragionevolmente la piccola massa dei quark e dei leptoni. L'argomentazione è la seguente: se i prequark sono particelle prive di massa, se hanno spin 1/2 e se interagiscono l'uno con l'altro solo mediante lo scambio di bosoni di gauge, è garantito che qualsiasi teoria che descrive il loro moto ha una simmetria chirale. Allora, se i prequark privi di massa si legano assieme per formare oggetti composti con spin 1/2 (cioè i quark e i leptoni), la simmetria chirale potrebbe garantire che anche le particelle composte rimangono prive di massa in confronto all'enorme energia dei prequark che sono al loro interno. La piccola massa dei quark e dei leptoni non è quindi un fatto casuale. Essi devono essere praticamente privi di massa rispetto all'energia dei loro costituenti, se deve essere mantenuta la simmetria chirale della teoria.

Il passo cruciale di questa argomentazione è l'estensione della simmetria chirale da un mondo di prequark privi di massa a un mondo fatto di quark e di leptoni composti. È essenziale che la simmetria del sistema fisico originario sopravviva e sia rispettata dagli stati composti formati dai costituenti privi di massa. Può apparire di per sé evidente che, se una teoria è simmetrica in qualche verso, i sistemi fisici descritti dalla teoria devono presentare tale simmetria; in effetti, invece, è comune la rottura spontanea delle simmetrie. Ne è un esempio ben noto la ruota della roulette. Una teoria fisica della ruota della roulette mostrerebbe che essa è completamente simmetrica nel senso che ogni settore è equivalente a qualsiasi altro. Il sistema fisico formato mettendo una pallina nella ruota della roulette è invece decisamente asimmetrico: la pallina finisce immancabilmente con il fermarsi in un solo settore.

Nel modello standard è la rottura spontanea di una simmetria a far sì che i tre bosoni deboli abbiano una grande massa e il fotone sia privo di massa. La teoria che descrive questi bosoni di gauge è simmetrica e in esse i quattro bosoni sono praticamente indistinguibili, ma a causa della rottura della simmetria gli stati fisici effettivamente osservati sono del tutto diversi. Le simmetrie chirali sono notoriamente sensibili alla rottura della simmetria. Se la simmetria chirale dei prequark si spezza o meno quando i prequark formano oggetti composti si può determinare soltanto da una dettagliata conoscenza delle forze agenti sui prequark. Finora tale conoscenza non esiste. In certi modelli si può dimostrare che esiste una simmetria chirale, ma che alla fine essa viene rotta. Nessuno è finora riuscito a costruire un modello composto di quark e di leptoni nel quale si sappia che una simmetria chirale è rimasta intatta. Neppure il modello a preoni né quello a rishoni riescono a risolvere il problema. Il compito è probabilmente il più difficile ostacolo per coloro che tentano di dimostrare che i quark e i leptoni sono composti.

Se si potesse elaborare una teoria coerente dei prequark, essa dovrebbe comunque superare la verifica sperimentale. In primo luogo, è importante stabilire in laboratorio se i quark e i leptoni abbiano o meno una struttura interna. In caso positivo, gli esperimenti potrebbero allora cominciare a discriminare tra i vari modelli. Gli esperimenti dovranno penetrare nel mondo sconosciuto di distanze inferiori a 10^{-16} centimetri e di energie superiori a 100 GeV. Vi sono due modi fondamentali per esplorare tale regione: eseguendo esperimenti con particelle accelerate ad altissime energie ed eseguendo misurazioni esatte di grandezze a bassa energia che dipendono dalla fisica di eventi a piccolissime distanze.

Esperimenti del primo tipo comprendono lo studio dei bosoni deboli e la ricerca delle particelle di Higgs del modello standard. Quando si potranno produrre tali particelle in numero sufficiente, un

La simmetria chirale offre una possibile spiegazione del «miracoloso» annullamento di massa nei quark e nei leptoni. La chiralità, o carattere destrorso o sinistrorso, descrive la relazione tra il momento angolare di spin di una particella e la sua direzione di moto. Supponiamo che un osservatore venga superato da un elettrone che si muove più rapidamente di lui *(a)*. Dal punto di vista dell'osservatore l'elettrone segue la regola della mano destra: quando le dita della mano destra si ripiegano nella stessa direzione dello spin, il pollice dà la direzione del moto. Però, se l'osservatore accelera in modo da superare l'elettrone *(b)*, la chiralità della particella cambia. Nel sistema di riferimento dell'osservatore l'elettrone si sta avvicinando anziché allontanarsi, ma la sua direzione di spin non è cambiata; di conseguenza il suo moto è descritto dalla regola della mano sinistra. La chiralità, perciò, non si conserva. Esiste un tipo di particella al quale non si possono applicare queste conclusioni, una particella priva di massa, che deve sempre muoversi alla velocità della luce. Nessun osservatore può muoversi più velocemente di una particella priva di massa e quindi la sua chiralità è una proprietà invariante. Se una teoria dei prequark avesse una simmetria chirale, nella quale si conserva il carattere destrorso o sinistrorso, quark e leptoni dovrebbero essere praticamente privi di massa per poter mantenere la simmetria chirale.

SIMMETRIA

ROTTURA SPONTANEA DELLA SIMMETRIA

La rottura spontanea della simmetria è un meccanismo che potrebbe vanificare una teoria dei prequark anche nell'ipotesi che essa possedesse una simmetria chirale. Entrambi i sistemi fisici mostrati in questo caso - un semplice ventre e un ventre con un rigonfiamento sul fondo - si possono descrivere come simmetrici nel senso che lo scambio della sinistra con la destra lascia il sistema inalterato. Per il ventre semplice il sistema rimane simmetrico se si mette una sferetta sul fondo; la sferetta si ferma al centro e uno scambio della sinistra con la destra continua a non avere effetti. Invece, nel ventre con rigonfiamento la sferetta prenderà posizione da una parte o dall'altra e la simmetria viene inevitabilmente spezzata. Allo stesso modo, una teoria dei prequark che presenta una simmetria chirale potrebbe dare origine a sistemi composti che non rispettano la simmetria. Il riuscire a dimostrare che una simmetria chirale può conservarsi indefinitamente è attualmente l'obiettivo principale nella formulazione di una teoria sul moto dei prequark.

accurato esame delle loro proprietà dovrebbe rivelare molte cose sulla fisica delle piccolissime distanze. Si prevede che nuovi acceleratori oggi in progetto o in costruzione negli Stati Uniti, in Europa e in Giappone forniranno dettagliate informazioni su bosoni deboli e porteranno avanti le ricerche in atto sugli stessi quark e leptoni.

Altrettanto interessanti sono gli esperimenti di alta precisione a bassa energia. Uno di essi è la ricerca del decadimento del protone, che è noto abbia una vita media di almeno 10^{30} anni. Numerosi esperimenti tengono oggi sotto controllo grandi quantità di materia comprendenti in pratica più di 10^{30} protoni, nel tentativo di rivelare i segnali emessi dalla disintegrazione di un protone. Nessuna delle forze del modello standard può indurre un tale evento, ma nessuna delle regole del modello standard lo vieta espressamente. Del resto, sia le teorie di grande unificazione sia i modelli a prequark comprendono meccanismi che potrebbero trasformare un protone in altre particelle che finirebbero col decadere lasciandosi dietro soltanto leptoni e fotoni. Se il decadimento verrà rivelato, la sua frequenza e la natura dei prodotti di decadimento potrebbero consentire di gettare uno sguardo determinante oltre il modello standard.

Il decadimento del protone è un evento ipotizzato che potrebbe essere interpretato come prova sperimentale per una teoria di grande unificazione o per un modello di sottostruttura dei quark. In un tipo di decadimento il protone si disintegrerebbe in un positone ($\bar{e}$) e in un pione neutro (π^0). L'evento si potrebbe spiegare in termini dei quark costituenti il protone: un'interazione dei due quark *u* trasforma uno di essi in un positone e l'altro in un antiquark $\bar{d}$; quest'ultimo si combina con il restante quark *d* del protone per formare il pione neutro. Le teorie di grande unificazione suggeriscono che l'interazione dei quark *u* è mediata da una nuova forza della natura. Il modello a rishoni fornisce una spiegazione alternativa: i due quark *u* si scambiano solo un rishone *T* e un rishone *V*.

Analogo interesse è destato dall'ipotetico processo nel quale un muone emette un fotone per essere quindi trasformato in un elettrone. Anche in questo caso nessuna delle forze del modello standard può portare a un evento del genere, ma anche questa volta nessun principio fondamentale lo vieta. Alcuni dei modelli composti consentono la transizione, altri no, di modo che una ricerca sul processo potrebbe offrire il modo di effettuare una scelta tra i modelli. Gli esperimenti eseguiti finora fissano un limite di meno di uno su 10 miliardi alla probabilità che un dato muone decada in questo modo. La rivelazione di tali eventi e la determinazione del loro tasso potrebbe gettare luce sulla misteriosa distinzione tra le generazioni.

Una terza classe di esperimenti di precisione sono quelli che continuano a perfezionare la misurazione del momento magnetico dell'elettrone e del muone. Sono prevedibili ulteriori miglioramenti sia nell'accuratezza sperimentale sia nei relativi calcoli dell'elettrodinamica quantistica. Se i risultati continueranno a essere in accordo con le previsioni del modello standard, il limite delle possibili dimensioni di una qualsiasi sottostruttura dei quark e dei leptoni apparirà più lontano. Se si scoprirà una discrepanza tra la teoria e gli esperimenti, essa costituirà un pesante indizio che quark e leptoni non sono elementari.

Potranno passare anche 10 o 20 anni prima di poter guardare chiaramente al successivo livello della struttura della materia (ammesso che esso esista). Ciò che si richiede è un solido quadro teorico, che sia coerente, che sia in accordo con tutti gli esperimenti e che sia abbastanza semplice da spiegare tutti gli aspetti del modello standard in termini di pochi principi e di poche particelle e forze fondamentali. Il quadro corretto, che si tratti di una teoria di grande unificazione o di un modello composto di quark e di leptoni, può già esistere in qualche forma embrionale. Del resto, è anche possibile che la teoria corretta emerga soltanto da qualche idea totalmente nuova. Nelle parole di Niels Bohr, può darsi che le nostre idee attuali «non siano abbastanza pazze da essere corrette».

Pavia e la cultura del Medioevo

Recenti ritrovamenti effettuati nell'Archivio di Stato pavese stanno gettando luce nuova sull'attività culturale della capitale del regno italico nel periodo medioevale

di Ugo Fiorina

Nel 1933 Max Planck scriveva che i metodi con cui la scienza opera nei vari campi rivelano stretta affinità e sembrano diversi solo per il loro vario adeguarsi all'oggetto preso in considerazione. In effetti ciò che rende unitario il criterio del sapere scientifico è il rigoroso confronto di un'esperienza, che può essere l'esame della traccia lasciata da una particella nel suo cammino in una camera a bolle o quello di un antico codice manoscritto, con altre analoghe esperienze precedenti per individuare i collegamenti utili a ordinare in una rappresentazione razionale la molteplicità del reale. E se per il fisico è essenziale garantirsi condizioni che non alterino i risultati dell'esperimento, altrettanto vale per lo storico che deve accertarsi in primo luogo della genuinità e veridicità delle fonti cui attinge i dati necessari al giudizio.

Nel campo delle scienze storiche lo sforzo di formulare criteri obiettivi per valutare l'attendibilità delle fonti librarie

I piatti di questo volume cinquecentesco, stampato dai Giunti Veneziani, dei «commentari» al *Digesto* e al *Codice* giustinianei del giurista pavese Giovanni Francesco Ripa sono stati rinforzati mediante incollatura con fogli mutilati a due codici manoscritti membranacei. I fogli nella parte bassa dell'illustrazione appartenevano a una *Eneide* del secolo XII. I versi visibili si riferiscono ai Canti XI (573-602), a destra, e XI (906-915) e XII (1-7), intervallati da monostico e decastico di Erennio, a sinistra. Il volume è il secondo dei due costituenti l'edizione del 1569 dell'opera, acquistata nel 1980 dall'Archivio di Stato di Pavia. In totale i due volumi recano incollati sui piatti 8 bifogli del codice virgiliano. I frammenti visibili presso il dorso vengono invece da un codice del secolo XIV.

Il chiostro quattrocentesco appartiene al monastero pavese di San Maiolo, dichiarato di interesse particolarmente importante dal Ministero per i beni culturali e ambientali. Il complesso, comprendente la chiesa, è il più antico e illustre insediamento cluniacense in Italia. La sua origine risale all'anno 967, ed è documentata da due carte coeve e dalla copia di un placito che ne confermano la destinazione a monastero maschile sottoposto a Cluny. Con l'abate Maiolo, che riforma San Salvatore a Pavia, Sant'Apollinare a Ravenna, San Paolo a Roma, e con il successore Odilone, il convento diventa la base dell'espansione dell'ordine cluniacense in Italia. Nel secolo XIV inizia la decadenza dell'antico priorato, che viene soppresso nel 1564. Alla fine del secolo XVIII monastero e chiesa sono venduti e profanati. Restaurato negli anni 1957-1960, il monastero è diventato sede dell'Archivio di Stato di Pavia, che spera di acquisire anche la ex-chiesa, fino a pochi mesi fa negozio di motocicli, ripristinando l'unità originaria del complesso, sorto sull'area di una delle antiche *insulae* della Pavia romana.

o documentarie e per separare quelle autentiche da quelle false ha portato alla creazione di discipline nuove che hanno, con il loro affinarsi, potentemente contribuito al progresso degli studi affiancandosi alla filologia. Le due più importanti, la paleografia e la diplomatica, sono nate insieme tre secoli fa nel trattato del monaco benedettino Jean Mabillon, pubblicato a Parigi nel 1681 con il titolo *De re diplomatica libri VI.*

Il Mabillon era stato spinto a scrivere da un preciso scopo: controbattere le accuse di falsità lanciate nel 1675 contro un gran numero di diplomi merovingi, conservati negli archivi dei monasteri francesi, dal padre bollandista Daniel van Papenbroeck nella introduzione al tomo II degli *Acta sanctorum Aprilis,* intitolata *Propylaeum antiquarium circa veri ac falsi discrimen in vetustis membranis.*

La confutazione, quasi un obbligo morale per i benedettini dal momento che i loro conventi serbavano la maggior quantità di carte dell'epoca merovingia, fu tanto valida da suscitare l'ammirazione dello stesso van Papenbroeck e da avviare ricerche simili anche nelle altre nazioni europee. Nel nostro paese i risultati critici delle indagini del Mabillon vennero messi a profitto dai giganteschi lavori di Ludovico Antonio Muratori.

Oggi la diplomatica, definita da Alessandro Pratesi, docente della Scuola speciale per bibliotecari e archivisti dell'Università di Roma, come «la scienza che ha per oggetto lo studio critico del documento al fine di determinarne il valore come testimonianza storica», rende servigi preziosi sia agli storici sia agli archivisti. Tra le conseguenze del perfezionamento di questa disciplina è da porsi la distinzione concettuale sempre più chiara tra due realtà solo apparentemente simili: l'archivio e la biblioteca. L'archivio, memoria scritta e ordinata di qualsiasi centro attivo organizzato, nasce con un preciso scopo pratico, quello di permettere alla persona o all'ente che produce i documenti di raggiungere determinate finalità. L'archivio conserva dunque documenti,

DE

RE DIPLOMATICA

LIBRI VI.

IN QUIBUS QUIDQUID AD VETERUM Inſtrumentorum antiquitatem, materiam, ſcripturam, & ſtilum; quidquid ad ſigilla, monogrammata, ſubſcriptiones, ac notas chronologicas; quidquid inde ad antiquariam, hiſtoricam, forenſemque diſciplinam pertinet, explicatur & illuſtratur.

ACCEDUNT

Commentarius de antiquis Regum Francorum Palatiis.

Veterum ſcripturarum varia ſpecimina, tabulis LX *comprehenſa.*

Nova ducentorum, & amplius, monumentorum collectio.

Operâ & ſtudio Domni JOHANNIS MABILLON, Presbyteri ac Monachi Ordinis S. Benedicti è Congregatione S. Mauri.

LUTECIÆ PARISIORUM,

Sumtibus LUDOVICI BILLAINE, in Palatio Regio.

M DC LXXXI.

CUM PRIVILEGIO REGIS, ET SUPERIORUM PERMISSU.

Questo è il frontespizio dell'opera che segnò la nascita di due nuove scienze, la paleografia e la diplomatica, fondamentali per la critica delle fonti storiche. Si tratta del *De re diplomatica libri VI*, trattato pubblicato a Parigi nel 1681 dal monaco benedettino Jean Mabillon.

ossia testimonianze scritte di fatti di natura giuridica, mentre la biblioteca è formata da manoscritti e libri redatti non per un fine immediatamente pratico, ma per ragioni culturali e di studio.

Questa distinzione sostanziale è oggi comunemente accettata e ne tiene conto anche la normativa che regola l'amministrazione dei beni archivistici e dei beni bibliografici nelle nazioni più evolute. L'Italia è il paese d'Europa più ricco di documenti di archivio d'interesse storico. Una indagine compiuta più di dieci anni or sono ha infatti rivelato che gli archivi di Stato italiani conservano più del doppio di tutta la documentazione storica reperibile negli altri stati europei assommati. Il valore pratico dei documenti prodotti nei secoli scorsi con il passare del tempo si è affievolito e a volte spento del tutto. Per esempio, gli ordini dati dal primo duca milanese Gian Galeazzo Visconti per l'edificazione della Certosa di Pavia hanno esaurito il loro scopo quando la costruzione è stata ultimata. Ma il valore di testimonianza storica di quei documenti è immenso per lo studioso di architettura, per il cultore di storia dell'arte, per lo storico delle finanze viscontee. Dagli ordini ducali e dalle conseguenti registrazioni contabili emergono i nomi degli architetti, le somme preventivate, i dati sulla manovalanza e sui materiali impiegati, sulle eventuali precedenti costruzioni abbattute, sui corsi d'acqua deviati. Quanto detto può naturalmente essere moltiplicato per moltissimi altri documenti di archivio, medioevali, moderni o contemporanei.

Con il decreto del Presidente della Repubblica del 13 dicembre 1975, n. 805, gli istituti statali italiani che amministrano il patrimonio archivistico prodotto e accumulato dai vari governi e istituzioni avvicendatisi nel tempo sono entrati a far parte del Ministero per i beni culturali e ambientali. La legge vigente prevede che ogni capoluogo di provincia abbia il proprio «Archivio di Stato», incaricato di incamerare mediante versamenti gli archivi anteriori all'ultimo quarantennio di tutti gli uffici statali (a eccezione di quelli dipendenti dal Ministero degli esteri e dal Ministero della difesa). Accanto a codesti archivi gli istituti possono anche accogliere, sistemare e offrire in consultazione agli studiosi gli archivi che enti pubblici o privati ritengano di voler promuovere a fini di ricerca scientifica.

La complessità della storia politico-istituzionale dell'Italia, nazione giunta assai più tardi di molte altre in Europa alla propria unità, nonché la ricchissima articolazione e fioritura di centri di potere (i Comuni), di cultura (le università), di natura ecclesiastico-religiosa (abbazie, monasteri, chiese), di assistenza (ospedali), hanno un puntuale riflesso nel numero sbalorditivo (per quantità e qualità) di archivi grandi e piccoli confluiti negli archivi di Stato. Se si potesse collocare tutta la documentazione (oltre un milione di pergamene e quasi otto milioni tra registri, cartelle e filze di atti cartacei) su un unico ideale palchetto, esso sarebbe lungo circa 950 chilometri. È quasi impossibile un giudizio di importanza poiché, se è vero che per le carte di natura politica e di governo gli archivi più ricchi sono quelli ubicati a Roma e nelle capitali pre-unitarie, per altre testimonianze di diversa natura (economica, sociale, culturale) l'archivio di Stato della più piccola provincia è sempre, per la storia di quel particolare territorio, il serbatoio più significativo di notizie.

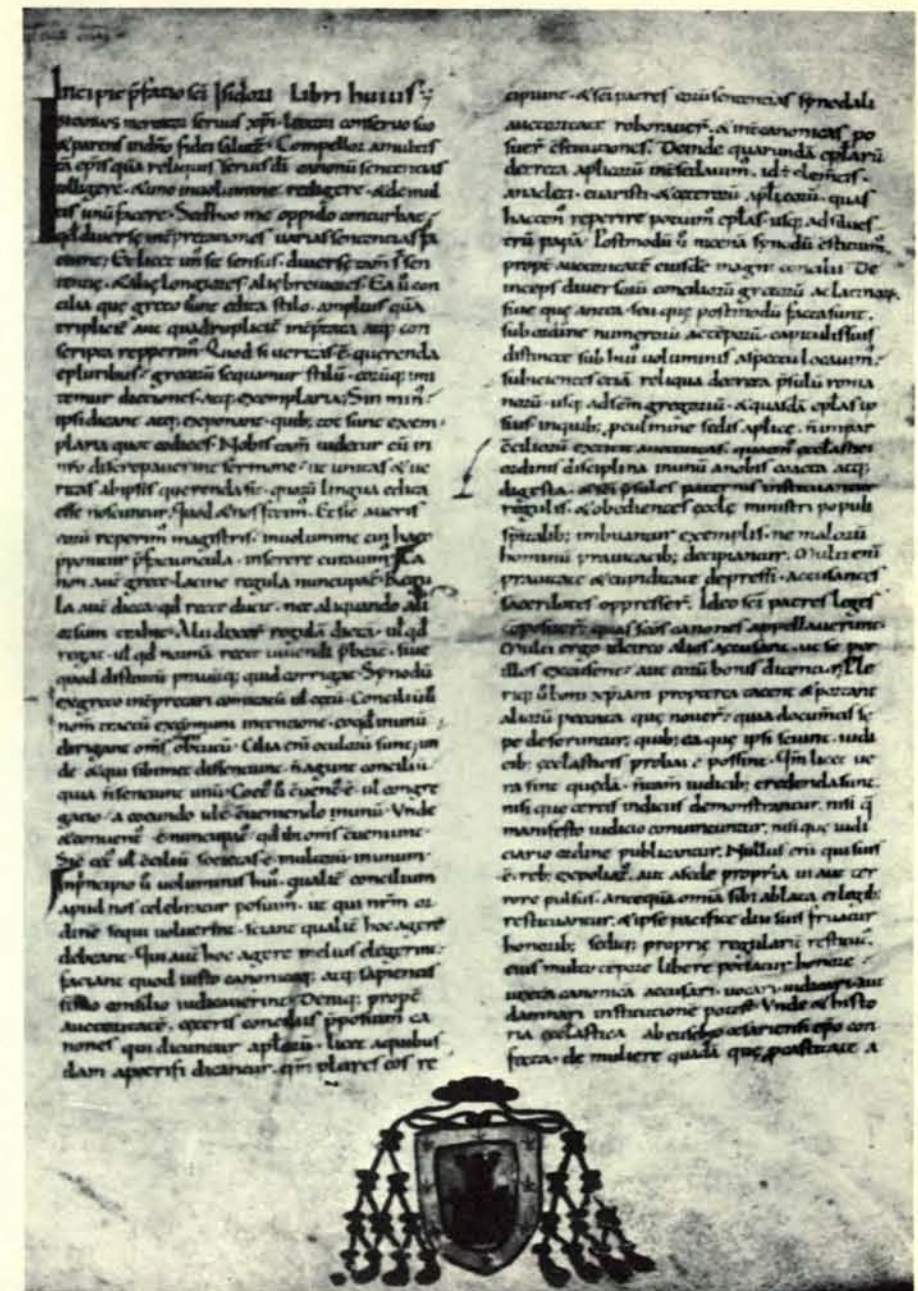
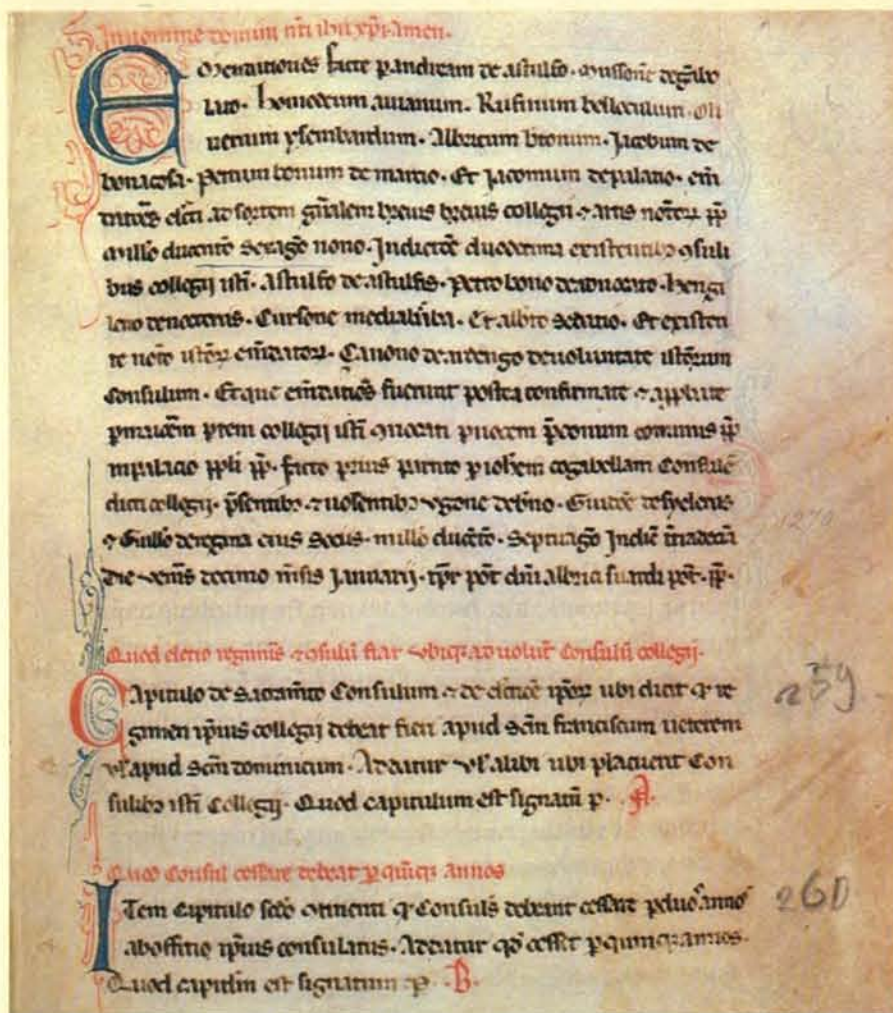
Per la genesi particolare del documento di archivio, che non nasce con la connotazione di universalità propria del libro, ma al contrario è strettamente radicato nel contesto politico-amministrativo locale, ciascun archivio di Stato è altamente individualizzato e i suoi fondi recano impressi nei dettagli i caratteri spirituali della collettività che, nelle varie strutture istituzionali, li ha prodotti. In qualsiasi archivio di Stato, almeno in Italia dove il notariato ha avuto un ruolo assai più significativo che altrove, non v'è fonte storica che superi, per vastità e ricchezza di contenuti, gli archivi notarili. Tutta la vita sociale, economica e anche culturale e artistica (i notai stendevano pure gli atti ufficiali delle università, i contratti di committenza degli artisti) di una città e del suo territorio è racchiusa nelle carte notarili, in gran parte non ancora studiate. Se nell'età romana classica i notai si configuravano come meri scrivani privati o addetti a pubbliche cancellerie, nel Medioevo (nel XII secolo) essi divengono pubblici funzionari con il potere di far fede dell'autenticità degli atti redatti. E a quest'epoca risalgono le imbreviature più antiche conservate in Italia, quelle del genovese Johannes Scriba, pur non mancando notizie sulle corporazioni anche altomedioevali dei notai, sull'istruzione che essi ricevevano, sulla loro attività. (L'imbreviatura era nel Medioevo la minuta dei negozi giuridici che per legge il notaio stendeva, di solito in forma abbreviata e faceva approvare dalle parti, riservandosi poi di redigerla in forma di istrumento.) Nell'Archivio di Stato di Pavia, che come gli altri omologhi istituti conserva gli archivi notarili del territorio (Pavia, Vigevano, Voghera) anteriori all'ultimo secolo, la documentazione prodotta dal Collegio e Arte (detta altrove anche *universitas* o *schola*) dei notai pavesi risale al secolo XIII e, tra le testimonianze più rilevanti che presenta, si trovano due statuti membranacei degli anni 1264 e 1266 e una matricola pure pergamenacea del 1284, con aggiunte posteriori.

A queste fonti sono da aggiungere gli atti veri e propri dei notai, che venivano di regola consegnati al Collegio alla morte del notaio stesso, e che per Pavia ammontano a 15 576 tra volumi, filze e registri per gli anni 1256-1870, con 429 repertori per il periodo 1400-1872.

I notai usavano anticamente riporre i propri atti in casse o armadi lignei, mentre il Collegio, stando ai più antichi inventari superstiti, sistemava la documentazione in casse. All'interno di queste ultime le imbreviature erano raccolte in pacchi di grandezza media,infilzati (da ciò il termine filza) da parte a parte con una cordicella munita di una punta metallica, per evitare dispersioni. Inoltre, perché la polvere non rovinasse le carte e per mantenerle ben diritte, il notaio, prima di trapassarle con il puntale, rinchiudeva il pacco tra due pesanti rettangoli di cartone, che proteggevano i documenti anche dal contatto diretto della cordicella avvolta strettamente intorno alla filza. Ai cartoni venivano preventivamente incollati per rinforzo fogli o frammenti minori di pergamena, tratti in gran parte da codici manoscritti e, in quantità più modesta, anche da documenti, breviari notarili, registri finanziari e testimonianze giudiziali.

La pergamena nel Medioevo aveva sostituito pressoché completamente il papiro come materiale scrittorio. Costituita da pelli di vitello, pecora e anche, per ottenere una qualità elevata, di agnellino, dal IV secolo era divenuta il tipico supporto dei codici librari, fino almeno al XV secolo, quando la carta, utilizzata in Italia a partire dal XII secolo, l'aveva rimpiazzata.

Nel prepararla per un documento la si lisciava e sbiancava solo dal lato interno (carne), poiché era l'unico a essere riempito dalla scrittura per intero. Se il documento era lungo, si cuciva o incollava al margine inferiore della pergamena un'altra membrana, e così via, formando in tal modo strisce lunghe anche diversi metri che si ripiegavano o arrotolavano. (L'atto membranaceo più antico conservato in Italia risale all'anno 716 ed è visibile presso l'Archivio di Stato di Milano.) Invece per i codici le pelli dovevano essere trattate accuratamente anche all'esterno (pelo), dato che si scriveva anche su questo lato.

Mentre per i documenti le pelli non ricevevano spesso una buona squadratura, nella costruzione del codice esse venivano ben rifilate per formare i fogli da piegare e riunire in «quaterni» (4 fogli doppi pari a 16 pagine attuali) o in rag-

La carta 1 recto del codice A 87 inf. della Biblioteca Ambrosiana di Milano contiene una redazione della prima metà del secolo XI del falso più illustre della storia. Si tratta della collezione di diritto canonico e romano denominata *Decretales Pseudo Isidoriane* dall'autore, un «Isidorus mercator». Composta assai abilmente in Francia nel periodo carolingio, contiene tra l'altro 50 canoni apostolici e 60 decretali pontificie falsi, nonché, in prosieguo, il testo pure apocrifo del famoso *Constitutum Constantini*, contenente la pretesa donazione dell'imperatore a papa Silvestro, fondamento giuridico, creduto autentico dagli stessi pontefici e come tale pubblicato ufficialmente, del dominio temporale dei papi. L'umanista Lorenzo Valla, che insegnò per qualche anno all'Università di Pavia, dimostrò la falsità del documento nello scritto *De falso credita et ementita Constantini donatione*, composto nel 1440. Questo e altri falsi nacquero in quell'epoca (secolo IX) dalla necessità in cui la Chiesa cattolica si trovava di arginare l'invadente ingerenza della dinastia franca nelle questioni religiose e dallo stato di decaduta efficienza degli archivi anche delle massime istituzioni (Impero, Chiesa), incapaci di organizzare un servizio che garantisse la conservazione e l'uso esclusivo di testi legislativi ufficialmente autentici.

gruppamenti diversi. Nei codici di buona qualità si formavano i raggruppamenti in modo che il lettore, aprendo il libro a qualsiasi foglio, avesse dinnanzi due facciate dello stesso colore, bianche (lato carne) o scure (lato pelo). L'operazione successiva era quella di segnare sui fogli gli spazi destinati alla scrittura, sia in senso verticale sia in senso orizzontale. Si cominciava con il fare una serie di forellini lungo i margini, per tracciare quindi le righe guida sulle quali si doveva scrivere. La griglia delimitava anche lateralmente il testo, che poteva essere a pagina intera o a colonne (due o più). Nell'alto Medioevo la rigatura era tracciata a secco, mentre a partire dal secolo XI sempre più spesso la troviamo fatta a matita. Sempre in quest'ultimo periodo si comincia ad annotare in fondo a ciascun raggruppamento di fogli (in genere un quaternio) la parola iniziale di quello seguente.

I sistemi usati per praticare i forellini e

La somiglianza tra queste due testimonianze scritte è solo apparente ed è data dalla comune scrittura. In realtà si tratta nel primo caso di un foglio di un libro manoscritto (seconda metà del XIII secolo) contenente biografie di imperatori romani, nel secondo *(in basso)* di un foglio di documento, lo statuto del Collegio e Arte dei notai di Pavia (1266). La differenza sostanziale tra biblioteca e archivio sta nel fatto che i libri della prima formano una raccolta di oggetti significanti anche indipendentemente l'uno dall'altro, mentre i documenti di un archivio costituiscono una *universitas* e sono tra loro strettamente interconnessi da un vincolo originario, come le tessere di un mosaico.

la rigatura variavano a seconda dei centri scrittori e rappresentano quindi un elemento significativo per l'attribuzione di provenienza dei codici. Quanto all'inchiostro, costituito principalmente da solfato di ferro e noce di galla disciolti, si nota che dal secolo VIII all'XI, contrariamente alle epoche precedente e seguente, esso in genere non è nero, ma color seppia scuro o rossastro.

L'amanuense altomedioevale tipico è il monaco. Il suo è un lavoro duro e interminabile. Non è raro trovare alla fine dei codici espressioni come «Chi non sa scrivere non può credere quanto sia penoso: ma sta il fatto che anche se sono solo tre dita a essere impegnate, il corpo intero si affatica». Se pensiamo che una Bibbia teneva occupato più di un anno un copista, non sorprende vederlo spesso esclamare, giunto all'ultimo foglio: «Il libro è terminato - Cristo sia lodato» o anche, meno compuntamente: «Il libro è finito - ora si dia da bere al copista». Grazie all'attività di migliaia di amanuensi medioevali, operanti presso le scuole capitolari, le abbazie, i monasteri più importanti, ci è pervenuta la maggior parte dei testi classici greci e latini. Il numero dei manoscritti per i vari secoli è in ragione inversa alla loro antichità. Essenziale fu il ruolo dell'ordine benedettino, il cui centro di Montecassino, fondato nel 529 da San Benedetto, diffuse in tutta l'Italia meridionale anche un nuovo tipo di scrittura, la beneventana. Ai monaci irlandesi, la cui figura più rappresentativa fu San Colombano, dobbiamo la istituzione, nei secoli VII e VIII, di una serie di monasteri che esercitarono un enorme influsso culturale nell'Europa continentale: Luxeuil e Corbie in Francia, Fulda in Germania, San Gallo in Svizzera, e il più illustre, Bobbio in Italia, la cui biblioteca è ora smembrata in tre nuclei all'Ambrosiana di Milano, alla Vaticana di Roma, e alla Nazionale di Torino.

Questo foglio di messale membranaceo notato ricopriva una volta un cartone di filza notarile, come si può vedere rovesciando l'illustrazione e leggendo in alto «MCCCCLXXXX SEXTO ET MCCCCLXXXX SEPTIMO / Filcia instrumentorum rogatorum per me Antonium de Clericis notarium publicum papiensem». Il frammento è stato scritto in minuscola carolina alla fine del secolo X, nell'Italia settentrionale, forse a Pavia. Si osservi la notazione musicale monodica del *Graduale* e del *Tratto*. La fissazione definitiva del repertorio musicale liturgico risale a San Gregorio Magno (604), ma i primi codici notati compaiono nel secolo IX. In essi le note, o i gruppi di note, sono costituite dai «neumi» (termine greco che significa «segno»), nati dalla trasformazione grafica degli accenti. Nel frammento i neumi sono posti in «campo aperto», senza indicazione di chiavi, di altezza tonale, di durata. Dal secolo XI, grazie alle innovazioni di Guido d'Arezzo, si cominceranno a usare i righi e le chiavi.

All'epoca della fondazione di Bobbio (612), Pavia è già da 40 anni sotto il dominio dei longobardi. Scelta come capitale del regno e sede perciò delle strutture centrali del governo politico e finanziario, con la conversione al cattolicesimo del rozzo popolo germanico invasore e la sua progressiva integrazione nel più elevato contesto sociale e civile dell'elemento romano indigeno guidato dalla Chiesa, Pavia si riempie di chiese, abbazie, conventi proporzionalmente anche alla sua aumentata importanza economica: San Salvatore, San Pietro in Ciel d'Oro, San Michele, Santa Maria alle Pertiche. I vescovati, le abbazie, i monasteri di maggior rilevanza (politica, economica e culturale) dell'alta Italia e anche d'oltralpe (San Gallo, San Martino di Tours), vi hanno le loro *cellae* o *curtes*, cioè residenze da usare nei momenti di necessità di presenza personale in occasioni di rilevanza politica (elezioni regie, assemblee) oppure giudiziaria.

Quando Carlo Magno espugna (774) la città, essa non perde il suo ruolo di capitale del regno italico, e beneficia dell'impulso innovatore carolingio conservando la sua posizione preminente fino al secolo XI. Nel 967 viene scelta dai cluniacensi come base per l'espansione dell'ordine e la connessa politica di riforma in Italia: nasce in quell'anno il convento di San Maiolo, attuale sede dell'Archivio di Stato. Tra gli altri, due celebri documenti del tempo dell'imperatore Lotario confermano il primato pavese. Il *pactum* dell'840 prevede tra i mercanti dell'impero carolingio e quelli di Venezia la reciproca esenzione da gravami fiscali e il libero commercio terrestre, fluviale e adriatico, sancendo la funzione del mercato e del porto di Pavia (sul fiume Ticino, vicino alla confluenza nel Po) nell'Italia settentrionale, come centro di collegamento e smistamento del traffico internazionale che congiunge l'Oriente, via Bisanzio-Venezia, con l'Occidente carolingio.

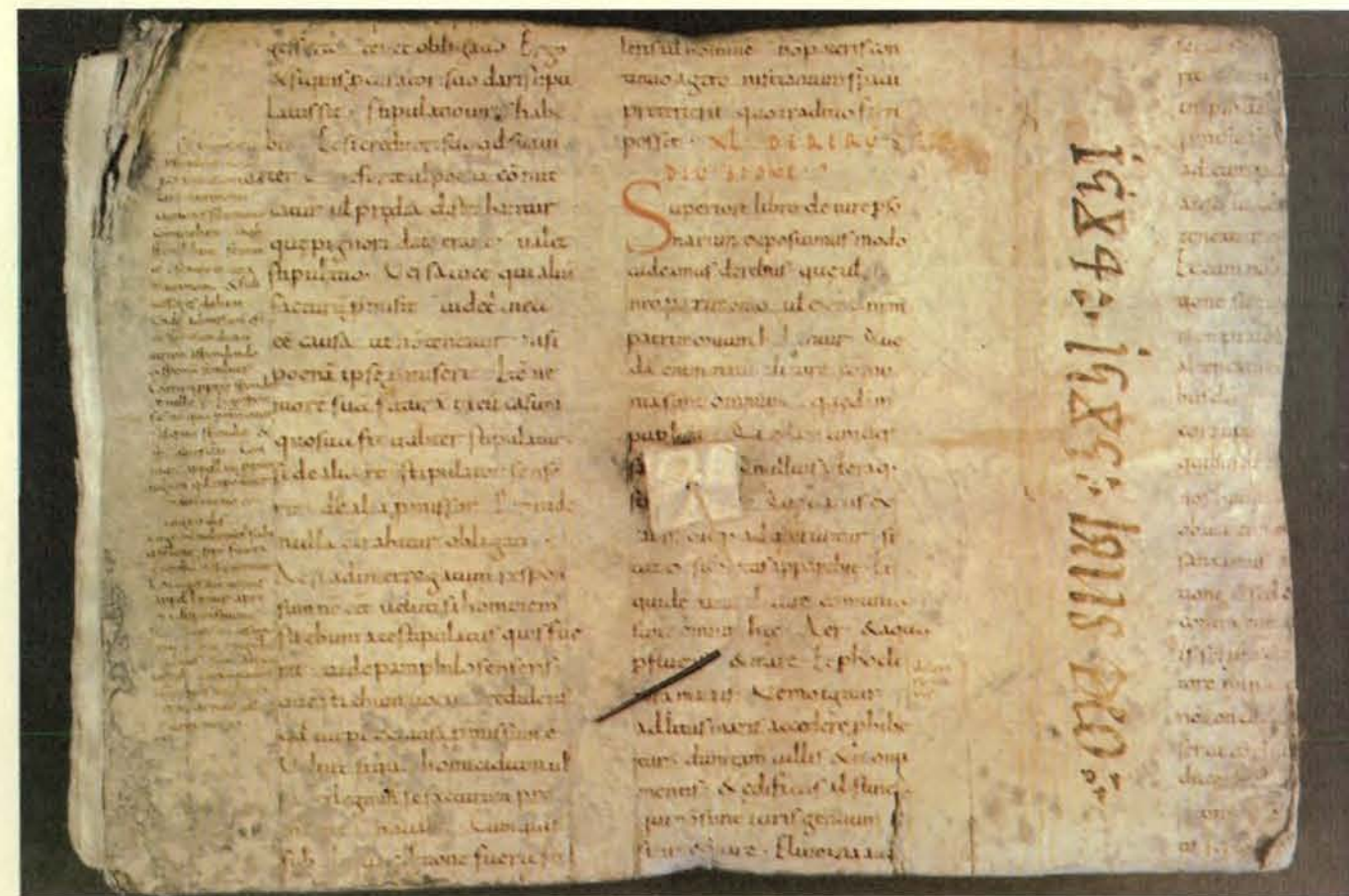
Quindici anni prima, l'attenzione di Lotario si era soffermata sulla necessità della riorganizzazione dell'istruzione

superiore nel regno italico. Con il capitolare dell'825, pubblicato a Corteolona (centro 15 chilometri a est di Pavia), egli dettava norme istitutive o confirmative di scuole in diverse città italiane. Pavia era la prima di esse, e dovevano convenirvi gli studenti di Milano, Brescia, Lodi, Bergamo, Novara, Vercelli, Tortona, Acqui, Genova, Asti e Como.

Il docente più illustre, e l'unico citato per nome nel documento, era Dungalo, monaco irlandese «*peregrinus pro amore Dei*» che a quel tempo si trovava a Pavia da oltre dieci anni, dopo aver insegnato nel monastero di St. Denis. Stimato da Alcuino e da Carlo Magno, che gli aveva rivolto richieste epistolari su questioni filosofiche e astronomiche, egli abitava ed esercitava nel pavese convento di Sant'Agostino, da dove poi si sarebbe trasferito a Bobbio, lasciando tutti i suoi codici all'abbazia.

Dungalo era il tipico esponente della cultura dell'epoca, poeta, teologo e amanuense provetto. L'insegnamento nell'epoca carolingia era svolto quasi esclusivamente dai monaci e dai chierici delle sedi vescovili. L'articolazione delle discipline, immutata per secoli, era quella, già attestata nelle *Etimologie* di Isidoro da Siviglia (m. 636), costituita dalle sette arti liberali e cioè grammatica, retorica, dialettica (trivio) e aritmetica, geometria, astronomia e musica (quadrivio).

A Pavia questa scuola, frequentata anche da Paolo Diacono, lo storico dei longobardi, non era la sola. Precise e inconfutabili testimonianze dimostrano che nella capitale del regno italico operava un centro di studi che non ebbe riscontro in nessun'altra città dell'Occidente altomedioevale. Codeste fonti sono rappresentate sia da testi legislativi, come il *Liber papiensis*, comprendente il corpus del diritto longobardo-franco vigente nel regno, sia da testi esegetici, di commento (glosse) a moltissimi capitoli del *Liber*, sia da formule. L'*Expositio longobardica*, un ampio commento, ci tramanda notizie abbondanti circa l'attività di diverse generazioni di giuristi (giudici e avvocati), che l'autore, ancora senza nome e vivente nel secolo XI, distingue in antichissime, antiche e moderne. Troviamo nomi illustri e ben noti anche alla posteriore scuola bolognese, come Bonfiglio, Walcauso e Lanfranco, che poi, passato in Normandia, diventerà abate del monastero di Bec, ed eleverà quel convento a un prestigio culturale di rilevanza europea; Lanfranco morirà in Inghilterra, arcivescovo di Canterbury (1089).

Nel lavoro di tutti questi giuristi cogliamo per la prima volta nella storia del diritto occidentale lo sforzo di sottomettere le norme barbariche dell'editto longobardo e dei capitolari franchi alla superiore validità del diritto romano: manifestazione giuridica di un travaglio di integrazione tra costumi violenti e una civiltà assai più alta, quella dei popoli assoggettati, che è la sostanza dei cosiddetti secoli bui del Medioevo. Si scontravano usi primitivi propri di genti germaniche, come nel diritto processuale quello del duello giudiziario, con consuetudini rispecchianti il connubio tra gli istituti romani e l'etica cristiana propugnata dalla Chiesa. Alla violenza, che improntava molti istituti germanici, si contrapponeva l'equità. Si riscoprivano i valori civili insiti nel diritto romano, riconosciuto allora per la prima volta «*lex omnium generalis*». Un fenomeno di tale portata non poteva manifestarsi che a Pavia, sede delle strutture centrali del regno, quali il *palatium* e la *curia palatina*, tribunale supremo.

La filza n. 3691 dell'archivio notarile di Pavia, contiene le imbreviature del notaio Giovanni Domenico Romussi quondam Giuliano. I documenti si riferiscono agli anni 1584 e 1585 e vennero racchiusi tra due rettangoli di cartone e infilzati da parte a parte con una cordicella munita di puntale. Il cartone anteriore reca incollato, come elemento protettivo, un foglio strappato a un codice membranaceo della fine del secolo IX o, al più tardi, del principio del secolo X. Si tratta del terzo foglio finora rinvenuto del più antico esemplare noto di una famosa collezione di diritto romano e canonico: la *Collectio Anselmo dedicata*. Il testo visibile è quello di una parte del II libro delle *Istituzioni* di Giustiniano. La glossa marginale, coeva, è tratta, come spessissimo accadeva nelle annotazioni giuridiche altomedioevali, dalle *Etimologie* di Isidoro da Siviglia. Diverse considerazioni fanno ritenere che la collezione, dedicata ad Anselmo vescovo di Milano nella seconda metà del secolo IX, sia stata composta da un ecclesiastico a Pavia, sede tra l'altro della più antica scuola giuridica medioevale attestata dalle fonti.

L'illustrazione mostra un frammento, ancora incollato al cartone, di un codice membranaceo del secolo XI contenente l'*Eneide* (VI, 895-901 e VII, 1-39 e 630-676). Si notano la rigatura a secco sul lato carne, una glossa nel margine superiore sinistro e il foro centrale per la corda della filza. I tre anni «1552. 1553. 1554», segnati da una mano coeva nel margine sinistro, indicano le date delle imbreviature che il notaio Giovanni Marco Maggi aveva protetto con il cartone. La scrittura è una minuscola carolina dell'Italia settentrionale. Le iniziali di verso, a colore alternato rosso e nero con tracce di argentatura, sono soprattutto di base onciale. Il frammento, un bifoglio quasi completo, è il secondo dei due rinvenuti dello stesso codice nell'Archivio di Stato e appartiene al più antico esemplare dell'*Eneide* mai trovato a Pavia.

Le arti liberali insegnate nell'ordinario corso di studi non potevano certamente essere sufficienti alla preparazione tecnica dei giudici chiamati a pronunciare sentenze o a occupare alte cariche, nonché dei vari «causidici», «*advocati*» e «*notarii sacri palatii*». Le esigenze poste da tale tirocinio richiedevano la presenza di una scuola, sempre beninteso nel senso altomedioevale del termine, e cioè raccolta intorno a una figura eminente. Per Bonfiglio, per esempio, i documenti la ricordano esplicitamente.

Le scuole non possono funzionare senza libri, e questi implicano l'esistenza di centri scrittori. San Pietro in Ciel d'Oro, San Salvatore, la curia vescovile, il *palatium regis* furono certamente a Pavia operosi *atelier* di codici. Purtroppo, mentre altri centri scrittori italiani e stranieri, come Montecassino, Verona, Bobbio, Corbie, Tours, Colonia, San Gallo, sono stati studiati nei loro particolari aspetti, quelli pavesi ancora attendono una ricostruzione scientifica che evidenzi per ciascuno le caratteristiche tipiche. Nel caso di Pavia la difficoltà maggiore è indubbiamente costituita dalla dispersione dei codici un tempo esistenti, ora sparsi nelle biblioteche europee e americane. Si pensi per esempio alla celeberrima biblioteca viscontea, ove lavorò il Petrarca, ora conservata alla Nazionale di Parigi.

I pochi codici integri rimasti a Pavia, presso la Biblioteca universitaria, non risalgono a prima del secolo XI, e rimane da stabilire se siano veramente stati scritti nella città. In mancanza di studi specialistici, si trovano spesso, per codici forse pavesi, attribuzioni di provenienza vaghe e prudenti che li assegnano a centri scrittori dell'Italia settentrionale non meglio identificati.

A nessuno sfuggirà il danno provocato da questo stato di cose agli storici delle varie discipline, dato il ruolo essenziale ricoperto da Pavia nell'età medioevale. Un contributo notevole e insperato sembra ora venire dal lavoro in corso di esecuzione presso l'Archivio di Stato di Pavia, diretto a recuperare tutti i cartoni delle filze dei notai pavesi, rinforzati, come si è visto, con frammenti di codici membranacei anche antichissimi.

Le 15 331 filze di imbreviature esistenti erano protette, quando pervennero all'Archivio, da due cartoni ciascuna, per un totale di 30 662 cartoni. È stata finora completata la ricognizione sistematica di 6801 filze, con il recupero di 2228 cartoni (non tutti i cartoni sono rinforzati). Se la media di un cartone ogni 3 filze è corretta, alla fine dello «scavo» si dovrebbe arrivare a circa 5000 cartoni.

Le dimensioni del rinvenimento, in termini di quantità e di qualità (i più antichi fogli completi risalgono al IX secolo, ma non si può escludere, per i motivi che si diranno, ritrovamenti di testimonianze più antiche), sono uniche in Italia. Infatti anche presso altri archivi di Stato (Genova, Savona, Siena, Parma, Cremona ecc.) si sono riscontrate analoghe presenze, ma non di pari consistenza e antichità. Per capire come ciò sia potuto accadere a Pavia, occorre ricordare quanto si è osservato circa l'eminente posizione rivestita dalla capitale del regno italico e l'esistenza di scuole e di ceti professionalmente legati, come oggi, alla necessità di libri (giudici, notai e anche chierici, monaci).

Per i notai, molte volte con certezza rei del «codicidio», occorre ricordare la tipica tendenza ad abbracciare l'arte paterna, ereditando con le imbreviature anche la biblioteca. E non vi è libro che invecchi

Angelo Matteo Bescapè, notaio pavese attivo dal 1596 al 1630, fece ricoprire tre suoi rogiti con altrettanti fogli tolti da un magnifico antifonario membranaceo del secolo XIV. Il lavoro in corso presso l'Archivio di Stato di Pavia ne ha permesso il recupero, che conferma ancora una volta l'influsso di Giotto sull'arte miniaturistica lombarda, come si può vedere nella figura del santo orante (San Sinforiano, nominato a cavallo tra la III e IV riga?) ritratta nella «G» iniziale di «Gloriosa».

più rapidamente, ai fini pratici, della compilazione giuridica, della collezione canonica: la nuova scienza bolognese ha provocato stragi dei testi precedenti, giustinianei e no. Discorsi simili si possono fare per i libri liturgici, abbandonati per le modificazioni imposte dalla Chiesa e per il perfezionamento dei sistemi di notazione musicale. La botta finale venne poi con l'invenzione e la diffusione, nella seconda metà del Quattrocento, dell'arte della stampa. Quando cominciarono a circolare le belle e chiare edizioni degli stampatori tedeschi, italiani, francesi, i libri manoscritti che fino ad allora avevano dominato il mercato uscirono dalla circolazione. Apparvero costosi (essendo scritto a mano il libro era caro: il valore medio oscillava dalle duecentomila lire di oggi in su e poteva valere milioni se era membranaceo, ben curato, rilegato e miniato), di non facile lettura, insomma sorpassati. Molti di essi, smembrati, finirono la loro carriera come rinforzi dei piatti e del dorso nelle legature dei trionfanti volumi a stampa.

Val la pena di accennare a come viene eseguito il recupero dei cartoni del notarile pavese. Si prendono in considerazione tutti i cartoni che si presentano rivestiti da frammenti, anche minimi, di pergamena o di carta, da un lato o dall'altro. Ciò è indispensabile perché i fogli o minori parti di rinforzo sono spesso stratificati, per addizione di nuove coperture alle precedenti, danneggiate o ritenute troppo deboli.

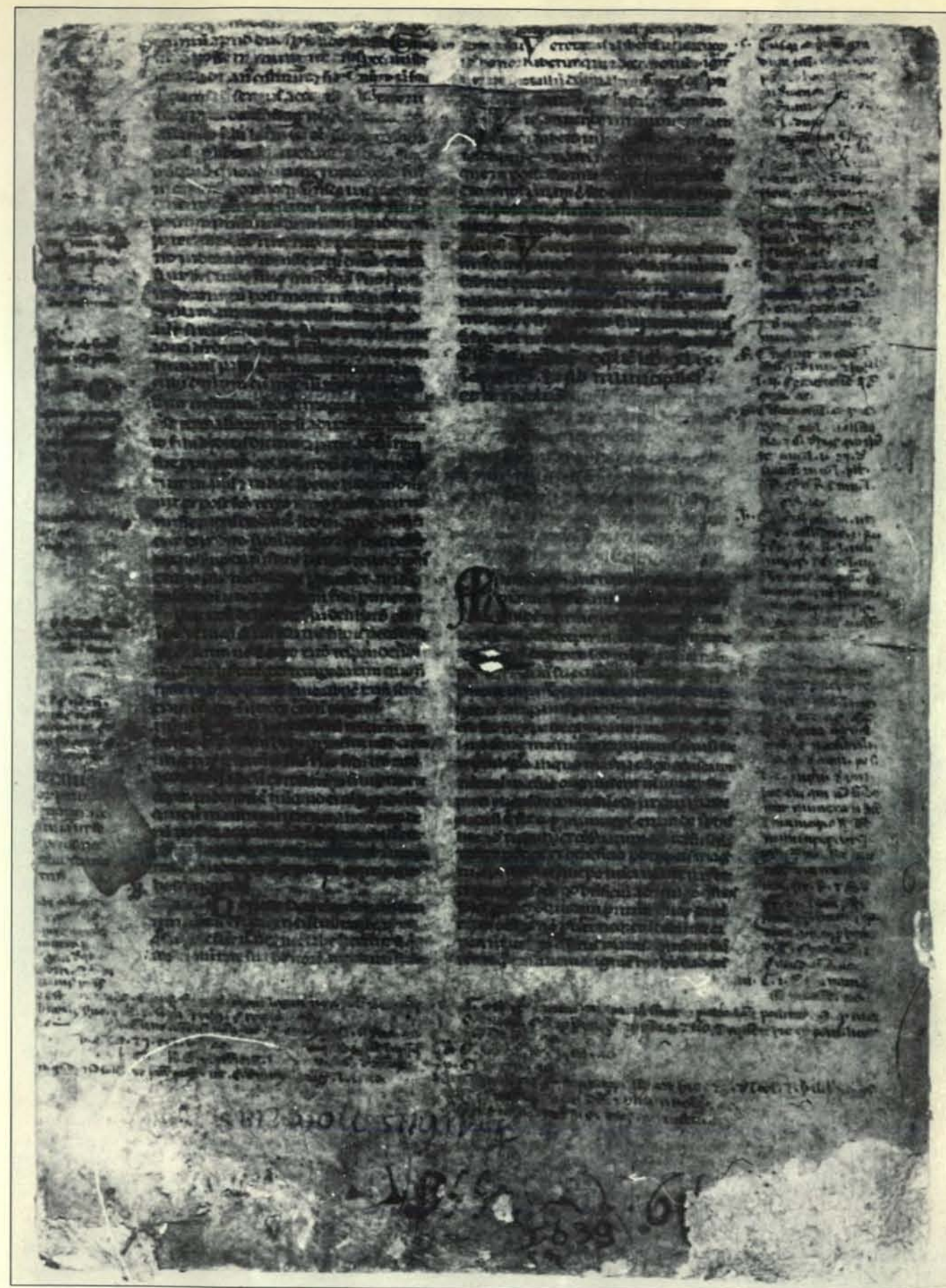
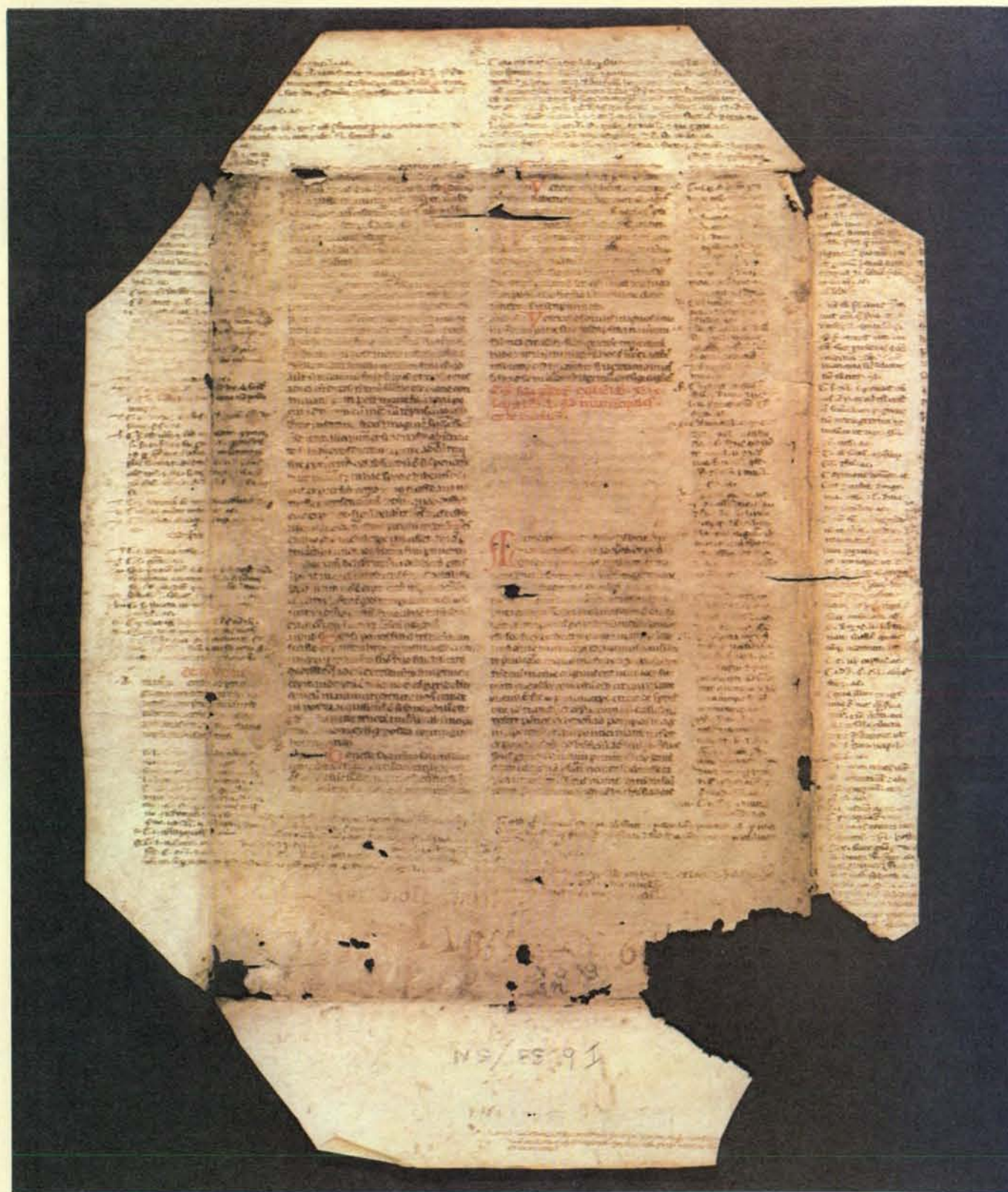
Si segna a matita sul cartone in esame, all'atto dell'estrazione della filza, il numero della filza stessa, in modo da ricollocarlo esattamente al posto originale qualora non interessi. Accertata la presenza di fogli o frammenti membranacei o cartacei scritti, o sospetti di esserlo (può trattarsi di fogli scritti solo sul lato interno, incollato), si recuperano quelli membranacei a qualsiasi epoca appartengano, e quelli cartacei fino al XVI secolo compreso, valutando quelli posteriori in base al contenuto.

Nel momento del recupero si compila per ciascun cartone una scheda con i seguenti dati: numero della filza notarile da cui proviene; nome, cognome e paternità del notaio che rogò le imbreviature ivi collocate; anni delle stesse; numero progressivo del cartone recuperato (giunto a tutt'oggi, come si è detto, al 2228). Contemporaneamente si segna quest'ultimo numero, a matita, sulla parte non rinforzata del cartone recuperato. Non si dà qui conto del lavoro analogo, in corso anche per i frammenti di codici o documenti usati per rinforzo o copertura degli atti delle filze vere e proprie.

Il cartone recuperato viene poi inserito in una copertina cartacea sulla quale si indicano datazione e contenuto dei frammenti, spesso di codici diversi.

Il contenuto è individuato con una lettera alfabetica di corrispondenza per una serie di categorie, fissate per comodità e quindi provvisorie: *B* (filosofia, storia, letteratura, grammatica, medicina, astrologia, geografia, storia naturale, dizionari, Antico e Nuovo Testamento); *C* (diritto civile e canonico); *E* (liturgia, musica sacra, notazioni neumatiche o quadrate); *F* (documenti pubblici e privati); *G* (codici ebraici). Le categorie *A* e *D*, inizialmente predisposte per i codici antichissimi *(A)* e quelli notati *(D)*, sono poi state rifuse nelle altre, la *D* passando integralmente nella *E*. Nella datazione dei documenti si è potuto, nei casi fortunati, individuare anno, mese e giorno. Quella dei frammenti dei codici librari è stata fissata con l'indicazione del secolo, sulla base di considerazioni paleografiche.

Nei suoi tre secoli di vita, la paleografia latina, inventata, come si è ricordato, dal Mabillon nel 1681, ha realizzato progressi giganteschi. Nell'opera del monaco benedettino essa aveva un aspetto classificatorio, che con le successive indagini (principalmente con L. Traube) si è arricchito dell'indispensabile interconnessione con la storia dei singoli centri scrittori e quindi della cultura nella sua accezione più vasta, nonché della consapevolezza dell'unità originaria della scrittura latina (già intuita nel Settecento da Scipione Maffei) e, in questo ambito, del chiarimento del rapporto tra le scritture librarie e quelle documentarie.

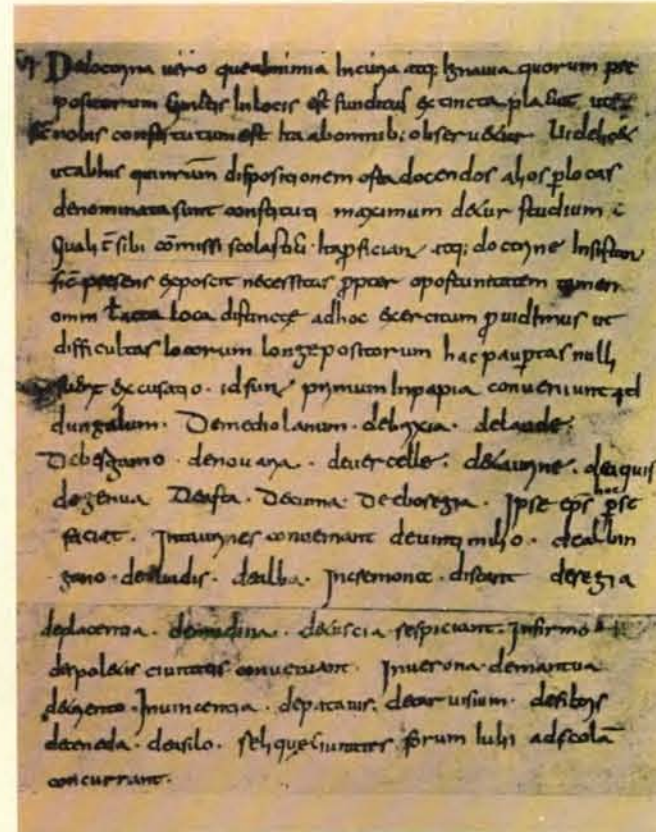
Librarie sono non solo le scritture dei libri manoscritti, ma quelle dei libri stampati. La forma dei caratteri di questa rivista, per esempio, è quasi invariata da 1000 anni. Infatti riproduce la scrittura che gli umanisti, e il Poggio (1380-1459) in particolare, veneravano e usavano chiamandola «*littera antiqua*», perché pensavano fosse quella dell'epoca romana. E la *littera antiqua* degli umanisti era esemplata sulla minuscola carolina evoluta del secolo XI. Anche i frammenti più antichi *(si veda per esempio l'illustrazione a pagina 66)* rinvenuti nel notarile sono scritti in carolina. Il nome deriva dall'es-

Un esempio di come si presentano i frammenti di codici membranacei prima e dopo il distacco. In questa pagina il frammento costituito, in questo caso, da un foglio di un *Digesto* con glossa accursiana del secolo XIV, è ancora incollato al cartone estratto da una filza di imbreviature cinquecentesche del notaio pavese Balbi. Sono evidenti i danni arrecati al testo dalla polvere e da vari altri agenti come la luce e gli sfregamenti con la filza vicina.

Qui sopra il foglio del *Digesto*, staccato dal cartone riprodotto nella pagina precedente, viene mostrato dallo stesso lato. La spianatura permette di vedere anche i margini, prima rimboccati e incollati dietro il cartone. Si possono osservare i tagli fatti agli angoli per avvolgere meglio il foglio al cartone e la miglior conservazione dei margini che, trovandosi all'interno della filza, si presentano assai più puliti e leggibili della parte centrale del testo, delimitata nettamente dalle antiche piegature. La fotografia di questa pagina mostra come il distacco consenta il recupero completo della facciata del foglio prima invisibile perché incollata al cartone. La metodologia, assai complessa, e il distacco sono stati direttamente elaborati dal Centro di fotoriproduzione, legatoria e restauro degli Archivi di Stato, che ha sede in Roma. Anche il materiale fotografico qui utilizzato è stato allestito dal Centro.

sere nata al tempo di Carlo Magno, contemporaneamente al rinascimento culturale promosso dall'imperatore e fiorito nelle scuole palatine e nei monasteri. Troviamo questa scrittura già pienamente diffusa nel secolo IX in Francia, Germania, Svizzera, Italia settentrionale e centrale. È scrittura ampiamente spaziata, dal tratto lineare, rotondeggiante e spontaneo. La sua bellezza sta soprattutto nella semplice essenzialità, secondo il canone che sarà poi caro a Giambattista Bodoni.

Con il passare del tempo queste caratteristiche subiscono alcuni mutamenti, importanti per l'assegnazione di un codice o di un frammento a questo o a quel secolo. Se si confrontano le illustrazioni che riportano esempi di scritture dal IX al XIII secolo *(si vedano le pagine 65, 66, 67, 73, 74 e 75)* si noterà come la carolina subisca, evolvendosi, un processo di compressione laterale, che fa sì che le singole lettere si allunghino verticalmente e si restringano. La semplicità originaria si perde nel fitto uso di trattini, le forme diventano angolose, dure, serrate. Lo spezzamento delle linee diritte, il tratteggio manierato e spesso pesante rivelano una nuova scrittura, figlia completamente diversa dall'armoniosa madre. È quella che gli umanisti chiameranno «gotica», cioè barbara, contrapponendole la loro «*antiqua*». La gotica è in realtà, come l'architettura, figlia dello spirito del suo tempo, lo spirito abbastanza intollerante e direi bellicoso delle grandi università (Parigi,

Questa splendida miniatura adorna un esemplare membranaceo tardoquattrocentesco degli statuti dell'Ospedale San Matteo di Pavia, fondato da una confraternita costituita da fra' Domenico da Catalogna e da 12 membri di famiglie nobili pavesi e istituito con bolla papale del 13 settembre 1449. La miniatura riempie la «S» iniziale del nome di papa Sisto IV, autore della bolla del 1479, premessa a questa redazione statutaria, e raffigurato in atto benedicente. Nel 1980 il San Matteo ha deciso di depositare presso l'Archivio di Stato di Pavia il proprio archivio, ricchissimo di testimonianze anche per la storia dell'abbazia di San Pietro in Ciel d'Oro.

Bologna, Oxford, ove si erano spostati l'insegnamento superiore e la ricerca scientifica, prima svolti presso i monasteri, le cattedrali, le scuole palatine), lo spirito della filosofia scolastica, con il suo rigore dialettico, l'ansia di non dimenticare, disputando, nessun dettaglio, e di inquadrare tuttavia ogni sottigliezza in uno schema rigido, intransigente. Una scrittura in armi, dunque, pronta a colpire, nata da una madre pacifica, serena. E per spiegarne l'origine penso si debba riflettere più sui cambiamenti culturali che non su fattori come il modo di temprare la penna o la posizione dell'amanuense rispetto al foglio da riempire.

In questo foglio, appartenente alla matricola membranacea del 1284, sono registrati i notai del rione di Porta del Ponte a Pavia i cui nomi iniziano con le lettere «a» e «b». Nel Duecento il prestigio del notaio tocca il suo apice: egli è per i suoi clienti, che possono essere sia semplici privati sia centri di potere politico o economico, il custode e garante dei negozi stipulati e l'interprete onesto e scrupoloso della loro volontà, attestata dall'imbreviatura o minuta del documento da lui conservato.

Con il capitolare olonnense dell'anno 825 Lotario detta norme per l'organizzazione dell'istruzione superiore nel regno italico. Nel capitolo VI, riprodotto dal codice 34 della Biblioteca capitolare d'Ivrea, si dispone tra l'altro che a Pavia, presso la scuola di Dungalo, convengano gli studenti di Milano, Brescia, Lodi, Bergamo, Novara, Vercelli, Tortona, Acqui, Genova, Asti e Como. Il prestigio dello studium pavese, soprattutto nelle discipline giuridiche, è attestato anche dall'*Expositio longobardica* del secolo XI.

Tornando al problema della datazione dei codici, è indubbio che il più lungo uso nel tempo di una scrittura accresce la difficoltà di collocare nell'epoca corretta un frammento magari di poche righe. La minuscola carolina fu adoperata dalla fine dell'VIII al XII secolo, quattrocento anni dunque. Qui preziosissimi sono per il ricercatore i criteri pratici offerti dalla scienza paleografica e accumulati, grazie alle osservazioni di generazioni di studiosi, mediante la comparazione di migliaia di codici, l'esame di singoli centri scrittori, il raffronto della produzione, anche miniaturistica, di questi ultimi.

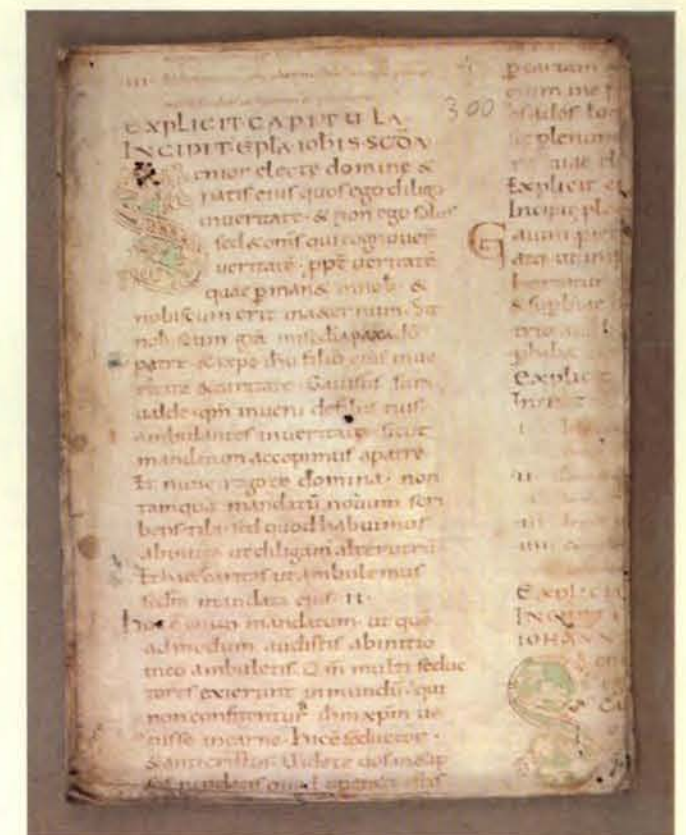
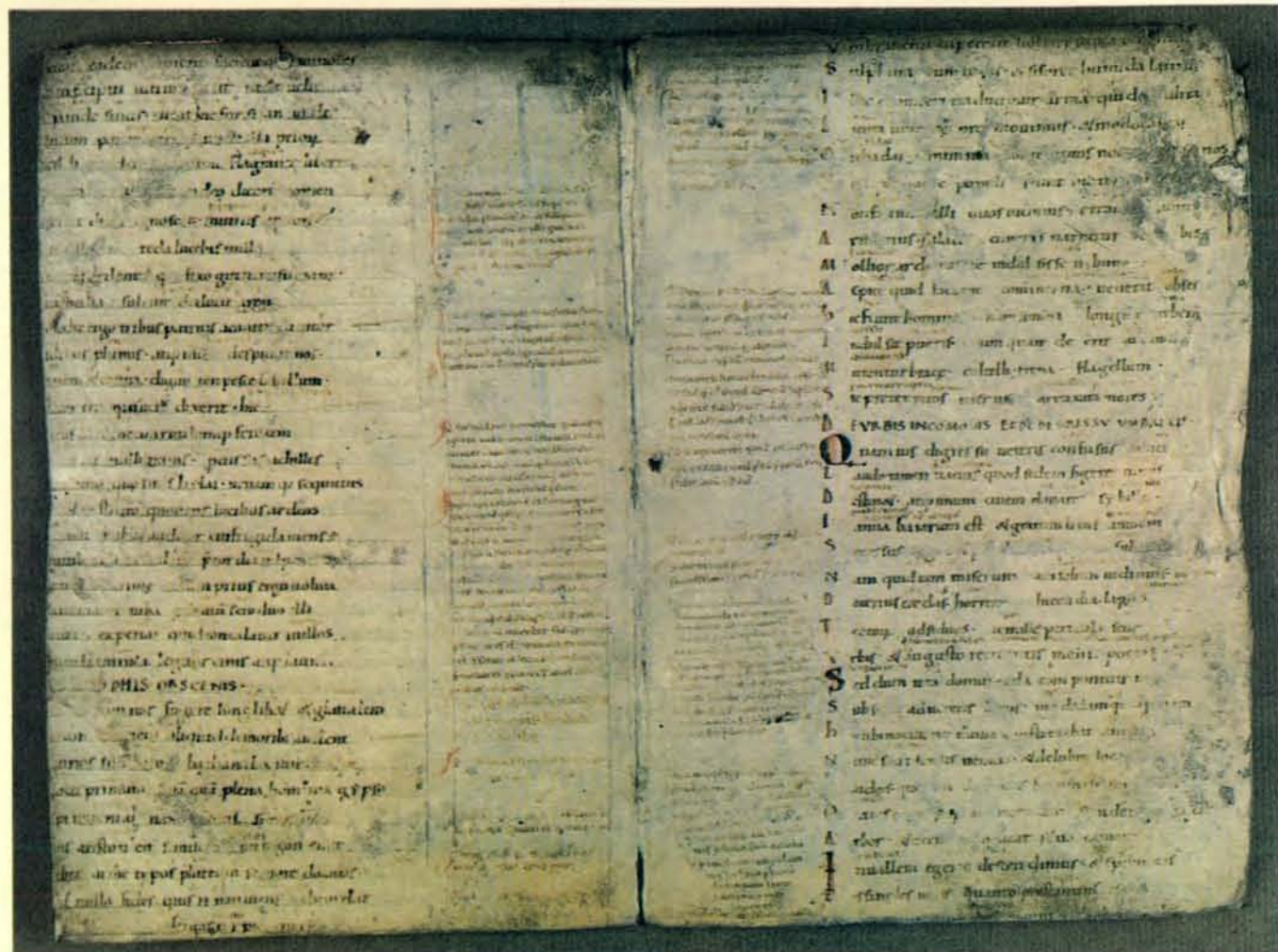
Esemplificando, si può dire, e il lettore potrà constatarlo sulle illustrazioni, che nel secolo IX la carolina presenta queste caratteristiche: la lettera «a» è di forma onciale (come la «a» di questa rivista) con la spalla destra assai inclinata, oppure di forma corsiva (come la «a» che adoperiamo nello scrivere) però generalmente aperta e cioè scritta come la «u». La lettera «g» di solito ha l'occhiello inferiore aperto. Le aste delle lettere «b, d, h, l» si ingrossano salendo, a forma di fuso. Il dittongo «ae» è raramente espresso con la forma ę.

Nel secolo X la «a» corsiva («u») si chiude e la spalla della «a» onciale comincia a raddrizzarsi. Si chiude anche l'occhiello della «g». Le aste delle «b, d, h, l» perdono gradualmente la forma a fuso. Si fa più frequente uso della forma ę per il dittongo «ae». Nell'XI secolo compare la forma della «s» rotonda, al posto della «s» diritta usata prima (come la «f» di questa rivista, senza il trattino trasversale); la «s» rotonda è però usata raramente e, in genere, in fine di riga e di parola. La spalla della «a» onciale è quasi sempre diritta. La forma «ę» è usata più della «ae» e spesso non correttamente. Le parole vengono ben separate l'una dall'altra, mentre prima erano fuse.

Con il XII secolo si accelera il processo di compressione laterale delle lettere, iniziato nel secolo precedente, con l'introduzione di trattini di coronamento. La forma «ę», ora la sola, alla fine del secolo scompare lasciando il posto alla semplice «e». Nello stesso periodo la «z» assume la forma «ç». La doppia «i» viene indicata con due segnetti diacritici che la distinguono dalla «u». La scrittura, diventata angolosa, zeppa di abbreviature, è ormai gotica.

Prima di concludere, vorrei dare un resoconto, sia pure sintetico e provvisorio, di quanto finora si è potuto identificare e datare nel corso del lavoro all'Archivio di Stato di Pavia. Per quanto riguarda i cartoni rinforzati con frammenti di incunaboli e cinquecentine, sono stati ritrovati 140 cartoni, nella categoria *B*, con frammenti di 61 libri, comprendenti edizioni di Aristotele, Boezio, Terenzio, Virgilio, Lucano, Galeno. Quattro cartoni portano frammenti membranacei dell'*Hypnerotomachia Poliphili* di Francesco Colonna, edita da Aldo nel 1499; 231 cartoni nella categoria *C*, per 65 libri di natura giuridica; 29 cartoni nella categoria *E*, per 13 libri liturgici, di cui tre in pergamena.

Per i cartoni rinforzati con frammenti di codici e documenti membranacei manoscritti, sono state trovate diverse centinaia di cartoni nella categoria *B*, con frammenti di codici glossati che vanno dal secolo IX al XV. Il contenuto copre filosofia (Aristotele, Cicerone, Boezio), poesia (Virgilio, Giovenale), teatro (Terenzio), agiografia (vite di San Martino, San Siro, San Lorenzo), grammatica, lessici (Papia), medicina, astrologia, storia naturale (Plinio), morale (San Gregorio Magno). Tra questi frammenti vi sono quelli dei codici più antichi di Virgilio e di Giovenale finora trovati a Pavia. Più di 200 i cartoni della categoria *C* con frammenti di oltre 160 codici glossati dal IX al XV secolo, relativi a diritto romano *(Corpus iuris civilis)*, diritto canonico *(Corpus iuris canonici)*, collezioni di canoni e decre-

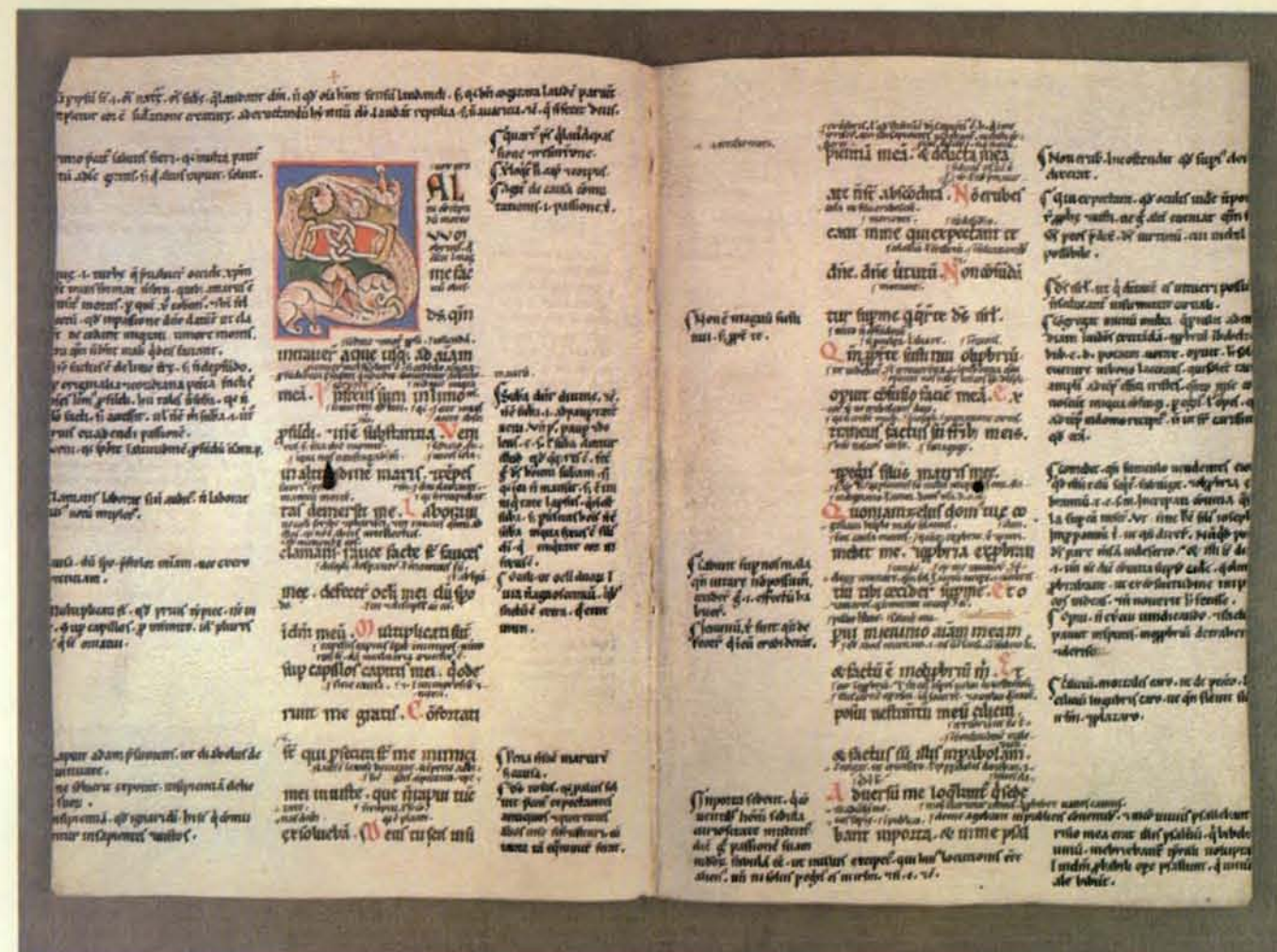
Nel frammento di questa *Passio beati Laurencii*, proveniente da un codice membranaceo del secolo XIII, la scrittura gotica è ormai completamente sviluppata. Il foglio è incollato a un cartone recuperato dalla filza n. 7087 del notaio Giovanni Battista Homodei, che rogò a Pavia dal 1596 al 1614. A Pavia esisteva anticamente una chiesa, oggi scomparsa, dedicata a San Lorenzo. Il culto di particolari santi, assai radicato localmente, è elemento prezioso per l'attribuzione delle fonti medioevali, spesso anonime.

Questo frammento membranaceo, ancora incollato al cartone di protezione della filza n. 2567 dell'archivio notarile pavese, appartenne a un codice biblico della seconda metà del secolo X. Si legge, in parte mutilata, la seconda lettera di San Giovanni. La scrittura è la minuscola carolina, tranne che per l'*explicit* e *incipit* dove è usata l'onciale. Le due «s» iniziali delle epistole sono colorate in rosso e verde e richiamano quelle del coevo *Sacramentario* di San Satiro, conservato a Milano.

Le fitte glosse marginali e interlineari rendono particolarmente significativo questo frammento, incollato su cartone, delle *Saturae* di Giovenale (I, 146-171 e II, 1-9 a sinistra; II, 155-170 e III, 1-19 a destra). Esso è, insieme ad altri tre dello stesso secolo XI, il più antico testo delle *Saturae* mai trovato a Pavia. Il cartone su cui è incollato ricopriva le imbreviature della filza n. 2869 del notaio Alessandro Folperti, che rogò dal 1539 al 1592. La scrittura del testo è la minuscola carolina. Le glosse sono di mani diverse dei secoli XI e XII.

tali, statuti e regole, trattati. Eccezionale è l'importanza dei tre frammenti della *Collectio Anselmo dedicata* che sono i più antichi finora noti della collezione altomedioevale. Nella categoria *E,* molte centinaia di cartoni, con frammenti di oltre 200 codici dal secolo IX al XVI: libri liturgici, sacramentari, innari, salteri ecc. I frammenti documentano oltre 60 codici notati. Oltre 200 i cartoni nella categoria *F,* con frammenti di documenti dal XII al XVI secolo: atti pubblici (bolle pontificie, concessioni sovrane ecc.), strumenti privati (imbreviature, registri contabili, investiture ecc.). Nella categoria *G,* 54 cartoni, con frammenti di 18 codici ebraici. L'esame compiuto da Pier Francesco Fumagalli, della Biblioteca Ambrosiana di Milano, dichiara il seguente contenuto: Bibbia, libri liturgici, poesia, un glossario ebraico-francese. Tutti i codici di provenienza sono del secolo XIV.

A tutte le unità sommariamente accennate sono poi da aggiungere un centinaio di cartoni i cui frammenti non sono per ora identificabili, sia perché logorati, sia perché incollati dalla parte scritta. Emerge da questa constatazione e da quanto già si è detto sulla stratificazione dei frammenti (un unico cartone porta talvolta frammenti di anche 6 diversi codici disparati per contenuto e data) l'esigenza sempre più sentita di un restauro totale dei reperti. Fino a quando tutti i cartoni non saranno restaurati, non sarà possibile il completamento di una delle operazioni più significative della ricerca: riavvicinare i frammenti di ogni singolo codice, incollati ora promiscuamente e a caso, spesso anche in *«sandwich»* formati da due cartoni farciti di fogli membranacei, ovviamente non esaminabili.

Sono frammenti di un universo, l'universo della attività culturale, istituzionale, giuridica, religiosa di uno dei centri più cospicui dell'Europa medioevale. Cultura è soprattutto relazione e perciò molti dei codici attestati non furono scritti a Pavia, ma a Bologna, Milano e altrove. Ma, anche nei codici materialmente non nati a Pavia, spesso troviamo glosse interlineari e marginali che dovranno essere attentamente considerate per la ricostruzione dell'ambiente scolastico pavese e di altri centri.

D'altra parte è evidente che solo il restauro consentirà di leggere le parti nascoste, che si sono tra l'altro meglio conservate, protette dalla polvere, dalla luce e dalle abrasioni causate dallo sfregamento con le filze contigue. Si tratta di un restauro assai delicato e complesso dato il valore delle testimonianze e i problemi posti dalla diversa composizione chimica delle colle usate nei vari periodi, degli inchiostri, dei colori e anche delle miniature che ornano spesso i testi. Anche la preparazione della pergamena e la sua qualità cambia a seconda dell'epoca e del centro scrittorio. Per questi motivi l'Archivio di Stato di Pavia si è rivolto al massimo istituto della tecnologia archivistica italiana, il Centro di fotoriproduzione, legatoria e restauro degli Archivi di Stato di Roma, che ha effettuato una ricerca e un diretto restauro su un primo gruppo di cartoni, rivestiti di frammenti di codici membranacei giuridici (dei secoli IX-XV) ed ebraici (del secolo XIV).

L'équipe, diretta da Vittoria Allemand Contini e formata da Cecilia Prosperi, Maria Teresa Tanasi, Paolo Ronci, Maria

La scrittura di questo frammento di una Bibbia membranacea (Salmo 68) mostra le caratteristiche della minuscola libraria verso la metà del secolo XII. La qualifica di carolina non sembra adattarsi più alle caratteristiche ormai gotiche (si notino le iniziali dei versetti) delle lettere. L'ampia spaziatura e la rotondità della carolina hanno lasciato il posto alle forme compresse e angolose della nuova scrittura. Il frammento è in questo caso un bifoglio mutilato usato come copertina di uno strumento del 1578, rogato dal notaio pavese Alessandro Folperti.

Molfese, Simonetta De Angelis, Massimo Santelli, ha messo a punto un nuovo procedimento che ha previsto, previa fotoriproduzione eseguita da Leopoldo Liccardo su pellicola ortocromatica dell'intero fondo, tre fasi: immersione del cartone in glicole etilenico (GE) fino a completo distacco della pergamena dal supporto; eliminazione del GE prima con carta da filtro, poi dalla pergamena con 3 bagni successivi di PEG 200 (polietilenglicole a peso molecolare medio 200) e dal cartone con tre bagni successivi di alcool etilico; eliminazione dalla pergamena dell'eccesso di PEG 200 con carta da filtro e successiva asciugatura sotto leggera pressione. Qualche piccola variante nel procedimento è stata resa necessaria per salvaguardare meglio i frammenti miniati.

Un risultato ulteriore del restauro, reso possibile dallo smontaggio delle stratificazioni plurisecolari avvolgenti i supporti, è quello della ipotesi formulata da Sergio Anzivino (del Centro) circa l'evoluzione storica delle coperture delle filze. Anzivino ritiene che in origine i cartoni fossero copertine di dimensioni più grandi, rivestite con manoscritti membranacei. Solo in un secondo tempo esse sarebbero state tagliate in formati minori e coperte per la seconda volta con altri frammenti membranacei, rimboccati sui piatti per proteggerne i margini tagliati. Infatti i frammenti più interni presentano, a seconda dei tagli effettuati, uno o più rimbocchi non combacianti con la parte aderente alla faccia opposta del supporto, appunto perché rifilati nella riduzione delle copertine. Circa le legature, la più recente è quella della filza con foro centrale; la più antica (copertina) doveva essere incartonata con nervature di pelle. Alcuni cartoni presentano ancora inseriti frammenti di pelle allumata, testimonianze delle precedenti e più grandi legature.

Il Centro, esorbitando dai compiti istituzionali, che sono quelli della sola ricerca e preparazione di nuove metodologie, ha direttamente restaurato 125 cartoni pavesi, recuperando 120 fogli interi e 50 frammenti di minor grandezza. Ora l'Archivio di Stato di Pavia dispone delle cognizioni tecniche per il restauro, ma non del laboratorio e degli operatori necessari. Occorre dunque affidare il lavoro a una seria ditta specializzata. Ma i mezzi finanziari sono venuti totalmente a mancare per i tagli nella spesa pubblica.

I cartoni da restaurare, oggi 2103, continuano ad aumentare di giorno in giorno. A questo punto varrebbe forse la pena chiedersi quando comincia a morire un libro. Mi pare che questa nota dimostri che un libro, come del resto le persone, muore quando non ci si interessa più di lui, quando non lo si ama più. Il restauro e lo studio dei frammenti pavesi diventano perciò il minimo dei doveri di ospitalità e di cortesia che abbiamo verso questi illustri viandanti, considerato il loro rango culturale e il lungo viaggio nel tempo da essi compiuto se vogliamo essere «savi e ben parlanti» come messer Torello di Stra quando accolse nella sua casa di Pavia il Saladino e i suoi compagni con le parole: «Signori, questa che voi ricevete da me, a rispetto di quella che vi converrebbe, per quello che io ne' vostri aspetti comprenda, fia povera cortesia; ma nel vero fuor di Pavia voi non potreste essere stati in luogo alcun che buon fosse, e per ciò non vi sia grave l'avere alquanto la via traversata, per un poco men disagio avere» (Boccaccio, *Decameron,* Giornata X, 9).

Le sorgenti calde sul fondo degli oceani

Scoperte lungo gli assi delle dorsali oceaniche, sono la manifestazione più spettacolare dell'attività vulcanica che, a grandi profondità, dà origine a nuova crosta e a ricchi giacimenti minerari sottomarini

di John M. Edmond e Karen Von Damm

Sul fondo dell'oceano, lungo le dorsali sottomarine, là dove aree estese di superficie terrestre vengono spinte lateralmente e allontanate e dove al loro posto sale nuova crosta oceanica, sono comuni le sorgenti calde. Il fenomeno avrebbe potuto essere previsto subito dopo la formulazione della teoria della tettonica a zolle, la quale spiega il significato delle dorsali, e in effetti l'ipotesi fu avanzata da J. W. Elder dell'Università di Manchester. Ora, dopo due decenni, la previsione trova conferma. Macchine da ripresa, sensori e sottomarini da ricerca che operano a profondità oceaniche prossime ai 3000 metri stanno individuando moltissime sorgenti calde e sembra che l'acqua che esse immettono in mare abbia rilevanti conseguenze.

In primo luogo, si sta dimostrando come il contenuto chimico dell'acqua termale, dovuto a una complessa serie di reazioni tra l'acqua marina e la roccia crostale calda di nuova formazione, serva a nutrire una catena di organismi viventi che va dai batteri ai bivalvi e a giganteschi vermi tubicoli. In secondo luogo, sembra che siano le reazioni idrotermali la fonte principale dei sedimenti e dei noduli ricchi in metalli che tappezzano il fondo dell'oceano. Lo sfruttamento di questi depositi è ora oggetto di discussione a livello mondiale. Si sta inoltre constatando come altri giacimenti minerari situati sui continenti e che sono stati sfruttati fin dall'antichità siano stati in realtà trasportati dove attualmente si trovano dai movimenti tettonici delle zolle dopo essersi formati sul fondo marino grazie all'attività idrotermale. In terzo luogo, sembra che il contenuto chimico dell'acqua termale sia determinante ai fini della composizione chimica degli oceani stessi.

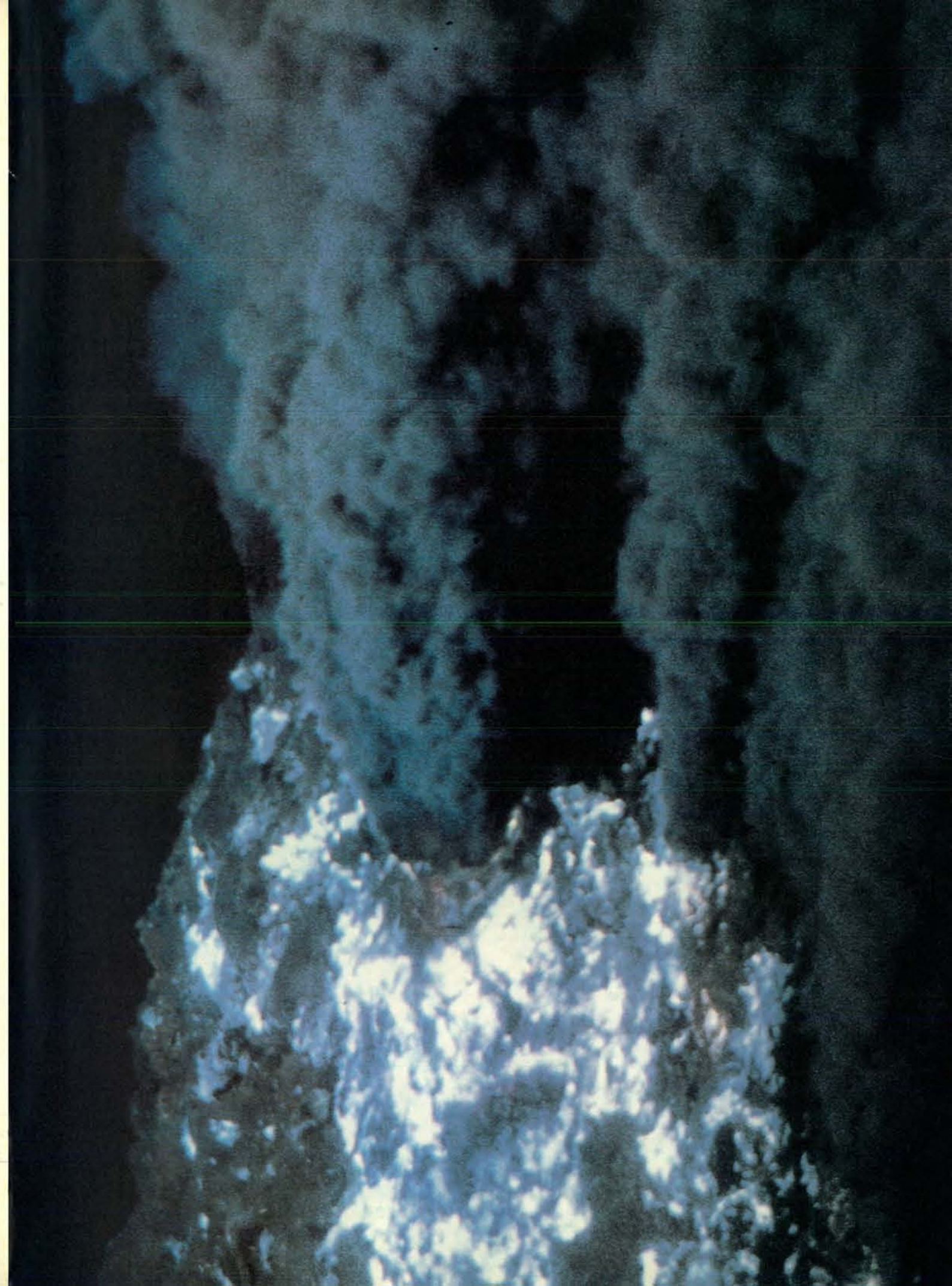
In retrospettiva, il posto delle sorgenti calde nella teoria della tettonica a zolle è inequivocabile. Secondo la teoria, la superficie della Terra è considerata come un insieme di estese zolle rigide in movimento costante l'una rispetto all'altra. Le zolle stesse sono per la maggior parte costituite da crosta sottostante gli oceani, con grandi «zattere» di materiale più leggero, i continenti, incastrate in esse. Ai margini dove le zolle collidono, viene distrutta la vecchia crosta. Questi margini sono fosse dove le zolle si immergono entro l'interno plastico del mantello. Ai margini dove le zolle divergono si forma invece nuova crosta. Questi margini si trovano quasi interamente sotto gli oceani e sono strutture lineari di un ordine di lunghezza di 100 chilometri, con una configurazione a zig zag determinata da faglie lunghe qualche chilometro.

Le dorsali medio-oceaniche

Lungo le strutture lineari, dove le zolle divergono, sale del magma (roccia fusa prodotta dalla parziale fusione del materiale del mantello a profondità non superiori ad alcune centinaia di chilometri), la cui temperatura al momento dell'arrivo è di circa 1200 gradi centigradi. Il magma si raffredda e solidifica dando origine a nuova crosta oceanica; poi, man mano che il raffreddamento continua, la crosta si frattura creando spazio per ulteriori intrusioni. Questo processo determina l'altezza del fondo oceanico. Il materiale caldo si espande e, quindi, le zone di intrusione sono elevate. Si tratta delle dorsali medio-oceaniche che si innalzano in mare a profondità comprese tra 2500 e 3000 metri. Raffreddandosi, il materiale si contrae e quindi il livello del fondo oceanico si abbassa man mano che ci si allontana dalle dorsali. Per questo motivo la maggior parte del fondo oceanico (e quindi la maggior parte della superficie terrestre) obbedisce a una correlazione semplice: la sua altezza diminuisce con la radice quadrata della sua età. Nell'Atlantico, la parte di fondo oceanico con un'età di 100 milioni di anni si è spostata a una velocità di circa un centimetro all'anno nella sua posizione attuale, ossia a circa 1000 chilometri di distanza dalla Dorsale medio-atlantica dove si è formata. Man mano che si raffreddava, si è abbassata di circa 2000 metri e così la profondità oceanica ha raggiunto i 5000 metri.

Poiché i centri tettonici di espansione delle dorsali medio-oceaniche sono zone di attività vulcanica, era facile immaginare come, alla stregua dello Yellowstone National Park e di altre aree continentali vulcanicamente attive, vi si sarebbero dovute trovare sorgenti calde, indice di attività idrotermale. Ma si aveva bisogno di prove e, ben presto, si sono avute.

Nel 1966 Kurt G. T. Boström e Melvin N. A. Peterson della Scripps Institution of Oceanography analizzarono campioni di sedimenti prelevati lateralmente agli assi delle dorsali medio-oceaniche. In questo modo constatarono come i centri di espansione oceanica individuati di recente fossero ricoperti da sedimenti ricchi di ossidi di ferro, manganese e altri metalli. Secondo questi ricercatori, i sedimenti erano dovuti alle reazioni idrotermali tra l'acqua marina e la calda crosta oceanica di recente formazione. Il lavoro successivo, condotto nell'ambito del Deep Sea Drilling Project,

La manifestazione più impressionante delle sorgenti calde e dell'attività idrotermale sul fondo dell'oceano è rappresentata dai cosiddetti «fumatori neri» che si trovano nelle dorsali medio-oceaniche. Quello che appare in figura è stato fotografato dal batiscafo da ricerca *Alvin* a una profondità di 2600 metri subito a sud dell'imboccatura del Golfo di California e si trova sulla cresta della Dorsale pacifica orientale. Ha un diametro di circa un metro e mezzo e l'acqua che emette ha una temperatura di 350 gradi centigradi. L'acqua permea la crosta oceanica, reagisce con la roccia basaltica calda di una zona vulcanicamente attiva al di sotto della dorsale e risale verso il fondo oceanico sotto forma di soluzione idrotermale: un liquido acido ricco di metalli. Al ritorno, la soluzione viene raffreddata dall'acqua marina e annerita dalla precipitazione di particelle di solfuro di ferro. Il camino del fumatore si forma anch'esso in seguito alla precipitazione dei minerali.

gestito dalla Scripps Institution, ha dimostrato che i sedimenti metalliferi sono sempre presenti alla base della colonna di sedimenti oceanici immediatamente sopra la roccia vulcanica della crosta oceanica.

Nel frattempo, la teoria della tettonica a zolle stava portando all'attenzione dei geologi le rocce della crosta oceanica. Questa è costituita da basalto, una roccia vulcanica nera. Chimicamente si tratta di un silicato di alluminio ($AlSiO_3$) privo di quarzo (SiO_2), ma ricco di ferro e di manganese. (Il granito e la riolite, invece, le rocce vulcaniche tipiche dei continenti, sono silicati di alluminio biancastri poveri di ferro e di manganese. In sostanza sono le «scorie» salite alla superficie dell'immenso forno di fusione che è la Terra.)

Negli anni sessanta vennero promosse campagne di dragaggio per consentire il prelievo di campioni da affioramenti superficiali nei pressi degli assi delle dorsali. Il Deep Sea Drilling Project recuperò dal fondo dei fori di trivellazione corte carote di roccia. L'analisi dimostrò che in generale questi campioni avevano subito almeno un episodio di reazione chimica con l'acqua marina dopo essere entrati a far parte della crosta oceanica. Per esempio, man mano che si raffreddavano fino a temperature inferiori, diciamo a 500 gradi centigradi, le rocce avrebbero potuto fratturarsi e l'acqua marina che penetrava nelle spaccature avrebbe potuto causare reazioni chimiche e riempire le fenditure di minerali precipitati, o secondari.

La conferma venne dai dati isotopici. In qualsiasi reazione chimica che coinvolga ossigeno, gli isotopi (ossia le specie nucleari) di questo elemento sono frazionati, cioè distribuiti non omogeneamente fra i prodotti di reazione. La ragione fondamentale è dovuta al fatto che l'energia vibrazionale di un legame chimico dipende dalla massa degli atomi legati insieme. Il grado di frazionamento è caratteristico di una particolare reazione e dipende inoltre dalla temperatura. L'analisi dei minerali secondari presenti nei campioni di basalto prelevati dai centri di espansione oceanica dimostrò, dunque, che la loro composizione isotopica di ossigeno differiva da quella del basalto e che i minerali si erano formati in seguito a reazioni del basalto con l'acqua marina a temperature di 350 gradi centigradi.

Le ofioliti

Ulteriori prove della presenza di attività idrotermale nei centri di espansione oceanica sono state raccolte inaspettatamente sui continenti. Si è visto che grandi lastroni di rocce basaltiche trovati sopra la crosta continentale nell'isola di Cipro, nel litorale mediterraneo nordorientale, nell'Oman, lungo il mare Arabico, in California, nel Tibet e altrove sono segmenti di crosta oceanica che sono stati spinti sopra i continenti dalle collisioni tra le zolle. Queste formazioni, chiamate ofioliti, hanno fornito l'opportunità di studiare in dettaglio i meccanismi di formazione del fondo oceanico. Nelle formazioni ofiolitiche era possibile campionare e cartografare facilmente ciò che una volta era stata crosta oceanica, spesso per centinaia di chilometri in senso orizzontale e per spessori anche di 10 000 metri. Anche in questo caso i segni di attività idrotermale ad alta temperatura furono impressionanti. Nella grande ofiolite dell'Oman, per esempio, è risultato chiaro che l'acqua marina era un tempo penetrata a profondità superiori ai 5000 metri e aveva reagito in maniera pervasiva con la roccia a temperature di circa 400 gradi centigradi.

Nella maggior parte delle ofioliti i basalti sono ricoperti da un sedimento metallifero, chiamato «terra d'ombra» (una varietà di ocra), spesso parecchi metri e identico a quelli che Boström e Peterson hanno trovato sul fondo oceanico attuale. Inoltre nelle ofioliti sono comuni i giacimenti minerari lenticolari contenenti milioni di tonnellate di solfuro di ferro (FeS_2, noto come pirite). Occupano depressioni sulla superficie del basalto e sotto di essi si trovano «camini» di quarzo e di minerali metalliferi. Chiaramente i camini erano una volta i condotti attraverso i quali salivano le soluzioni idrotermali da cui sono precipitati il quarzo, i minerali metalliferi e i giacimenti minerari lenticolari.

Questi bivalvi giganti, fotografati sul fondo oceanico in un campo di sorgenti calde a sud del Golfo di California, si trovano tra cuscini di basalto lungo le fenditure da cui sgorgano soluzioni idrotermali la cui temperatura è di circa 17 gradi centigradi. Le soluzioni sono ricche di acido solfidrico di cui si nutrono certi batteri, i quali, a loro volta, nutrono questi bivalvi.

Cercando di determinare in maniera diretta l'equilibrio termico della crosta oceanica man mano che essa si evolve, in tutti gli oceani sono state effettuate misurazioni della perdita di calore per conduzione dal fondo oceanico. La perdita per conduzione rappresenta il flusso di calore attraverso un mezzo immobile; in questo caso, i sedimenti che coprono il fondo oceanico. In generale risulta che la perdita per conduzione diminuisce andando verso le dorsali medio-oceaniche. Eppure, nelle dorsali la crosta dovrebbe essere più calda della crosta più antica più distante. Gran parte, quindi, del calore portato alle dorsali dal magma che sale deve essere rimosso da un altro meccanismo, la convezione, ossia il trasporto di calore da parte di un fluido che, in questo caso, è l'acqua marina. In effetti, Clive R. B. Lister dell'Università di Washington ha accertato che i processi termali lungo l'asse delle dorsali devono essere dominati dalla convezione.

L'integrazione su scala globale della differenza tra la perdita di calore misurata e la perdita che ci si aspetterebbe se questa avvenisse interamente per conduzione dà un valore di circa 5×10^{19} calorie per anno, ossia circa un decimo del flusso totale di calore proveniente dall'interno della Terra. Questo valore si deve alla convezione. Supponiamo che la temperatura di regime delle celle convettive sia di 350 gradi centigradi, in accordo con i dati isotopici; allora un volume di acqua marina uguale al volume degli oceani ($1,37 \times 10^{21}$ litri) deve circolare attraverso gli assi delle dorsali e, quindi, a un regime di temperatura di 350 gradi centigradi, ogni otto milioni di anni circa. È un tasso pressappoco pari allo 0,50 per cento del tasso al quale gli oceani ricevono acqua dai fiumi della Terra.

Gli isotopi dell'elio

Forse la prova più convincente dell'importanza attuale dell'attività idrotermale negli oceani deriva dalle misurazioni della concentrazione di elio nella colonna d'acqua oceanica. L'elio ha due isotopi, l'elio 3 e l'elio 4. Nel momento in cui la Terra si è formata, i due isotopi avevano un certo rapporto reciproco di abbondanza. L'elio 4, però, viene prodotto all'interno della Terra dal decadimento radioattivo degli isotopi a lunga vita dell'uranio e del torio. Inoltre, sia l'elio 3 sia l'elio 4 sfuggono nello spazio interplanetario dalla parte alta dell'atmosfera terrestre e, allo stesso tempo, vengono entrambi immessi nell'atmosfera dal vento di particelle che, proveniente dal Sole, investe la Terra.

I rapporti tra elio 3 ed elio 4 nelle varie località della Terra sono quindi il risultato di vari processi. Brian Clarke della McMaster University ha messo a punto un sensibile spettrometro di massa che misura con grande precisione la composizione isotopica dell'elio e lo ha impiegato per analizzare l'elio disciolto nell'acqua che fluisce nei pozzi dalle falde acquifere. Si suppone che, dal momento in cui l'acqua è penetrata inizialmente nella falda, la concentrazione di elio 3 non abbia subito alcuna variazione e quindi rappresenti la solubilità di equilibrio dell'elio, ossia la quantità di elio che passa dall'atmosfera all'acqua per via naturale. Qualsiasi anomalia nella concentrazione di elio 4 rispetto al valore di equilibrio si può attribuire al decadimento radioattivo sotterraneo. In poche parole, lo spettrometro è utile nella prospezione dell'uranio.

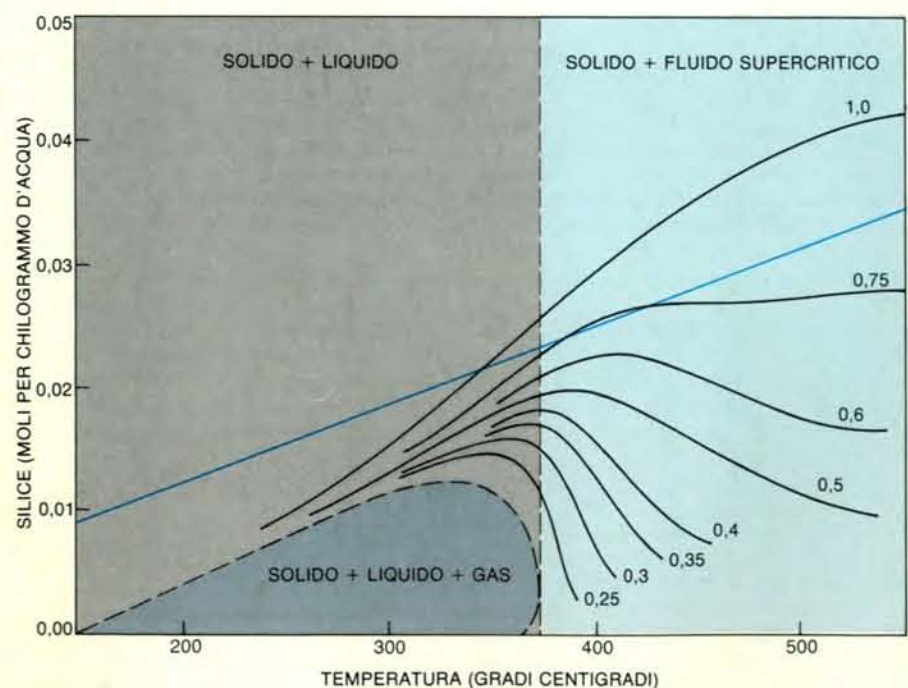
Stabilito questo, Clarke, in collaborazione con Harmon Craig della Scripps, ha tentato di fare una stima della quantità di elio 4 che passa dal mantello nell'oceano. Come nella valutazione precedente, l'elio 3 fu preso come indice, ma ciò che scoprirono fu notevole. Nel Pacifico, l'elio disciolto nello strato superficiale dell'acqua era effettivamente in equilibrio con l'elio presente nell'atmosfera. A profondità maggiori la quantità di elio superava il valore di equilibrio e, quando quest'ultimo veniva sottratto dalla concentrazione misurata, la quantità di elio 3 in eccesso era anormalmente grande. La differenza raggiungeva il massimo a 2600 metri, ossia alla profondità media dell'asse della dorsale del Pacifico. Nell'elio in eccesso il rapporto tra elio 3 ed elio 4 era circa otto volte il valore atmosferico. È evidente che l'elio 3 è contenuto nel mantello e viene liberato nell'oceano in corrispondenza degli assi delle dorsali. L'elio sfugge semplicemente da bolle di gas presenti sulla superficie del magma che si raffredda, oppure viene liberato nel corso dell'attività idrotermale che trasforma le rocce? La scala della differenza fa propendere per la seconda ipotesi.

La composizione dell'acqua marina

In che modo esattamente l'acqua marina reagisce con il basalto a pressioni che, a livello del fondo oceanico, sono di circa 300 chilogrammi per centimetro quadrato e a temperature di centinaia di gradi centigradi? Le prime risposte si ebbero verso la metà degli anni settanta quando James L. Bischoff e Frank W. Dickson nel loro laboratorio alla Stanford University riprodussero quelle condizioni. I risultati furono impressionanti. Gli ioni magnesio nell'acqua marina si combinarono con i silicati presenti nel basalto, formando un idrossisilicato, $Mg(OH)SiO_3$, insolubile in acqua. I gruppi ossidrile (OH) vennero ceduti dall'acqua, dove rimasero ioni idrogeno (H^+); per questo motivo la soluzione di acqua marina diventò fortemente acida. Gli ioni idrogeno poi presero il posto del calcio e del potassio nel reticolo cristallino del basalto e così il reticolo cristallizzò di nuovo.

In sostanza le reazioni consistettero in uno scambio di ioni idrogeno con altri

cationi, ossia ioni positivi, e in particolare con ioni calcio e potassio. Il calcio liberato poté così reagire con il solfato (SO_4) presente nell'acqua marina e formare solfato di calcio ($CaSO_4$), il quale precipita in forma di anidrite. Il solfato poté anche reagire con il ferro presente nel basalto e dar luogo a solfuro di ferro, ossia a depositi di pirite. Queste due reazioni provocarono la totale rimozione del solfato dal-

Il primo segno che l'attività idrotermale doveva costituire un'importante caratteristica delle dorsali medio-oceaniche è stato il flusso di calore proveniente dal fondo oceanico. Nelle dorsali, zolle di crosta oceanica si allontanano l'una dall'altra lasciando il posto alla risalita di nuova crosta. Per questo motivo le dorsali sono molto calde. Ciononostante la conduzione termica attraverso la crosta misurata in tre dorsali atlantiche *(in colore)* era risultata molto inferiore al previsto nell'ipotesi che tutto il flusso avvenisse per conduzione *(in nero)*. La differenza va attribuita alla convezione, ossia all'acqua che passa attraverso il fondo oceanico. L'asse orizzontale indica l'età della crosta oceanica, un parametro della sua distanza dalla dorsale dove si è formata.

La solubilità del quarzo nei campioni di acqua termale prelevati nel 1977 dai campi termali sul fondo oceanico nei pressi delle isole Galápagos ha fatto pensare, ancor prima della scoperta dei fumatori neri, avvenuta due anni dopo, che le soluzioni idrotermali pure che risalgono verso il fondo oceanico dovessero avere una temperatura di circa 350 gradi centigradi. La temperatura dei campioni prelevati alle Galápagos non superava i 19 gradi centigradi, ma la concentrazione di silice (quarzo disciolto) era massima nei campioni più caldi. L'estrapolazione della tendenza *(in colore)* intersecava le curve relative all'acqua satura di quarzo *(in nero)* in punti molto al di sopra dei 300 gradi centigradi. (Le pressioni sono indicate in chilobar.) Era evidente che i campioni delle Galápagos rappresentavano soluzioni idrotermali diluite con acqua marina comune. Il grafico mostra i regimi in cui una miscela di quarzo e acqua è stabile come (1) solido più liquido più gas, (2) solido più liquido e (3) solido più fluido supercritico (una fase che non è né liquida né gassosa).

l'acqua marina. Infine, la degradazione del basalto e l'acidità della soluzione di acqua marina consentirono la liberazione di silice (SiO_2, disciolta in acqua).

La composizione chimica degli oceani potrebbe essere dovuta tanto a queste reazioni idrotermali quanto ai prodotti della disgregazione causata dagli agenti atmosferici sui continenti? L'eventualità era particolarmente interessante perché a metà degli anni settanta la comprensione dei processi che determinano la composizione dell'acqua marina si trovava in un vicolo cieco. Nel XIX secolo, prima del grande sviluppo della chimica organica, molti chimici si erano dedicati all'analisi di sostanze naturali inorganiche.

Questo indirizzo era motivato soprattutto dalla ricerca di elementi chimici ancora sconosciuti che colmassero i vuoti della tavola periodica degli elementi, da poco elaborata. La principale strategia seguita consisteva nell'analisi di minerali metalliferi e R. W. Bunsen, per esempio, ne analizzò centinaia. Si interessò così dei processi magmatici e si spinse fino in Islanda per osservare direttamente le eruzioni vulcaniche. Altri ricercatori si dedicarono allo studio delle acque naturali e fu così che J. H. van't Hoff gettò le basi della geochimica sperimentale studiando la sequenza dei precipitati che si formano quando l'acqua marina evapora.

Grazie a tutte queste ricerche fu chiaro, alla fine del secolo, che l'acqua marina non poteva essere il risultato della parziale evaporazione dell'acqua dei fiumi. Alla fine di questo processo possono esservi solo bacini chiusi come il Mar Morto o il Great Salt Lake, i quali, in confronto agli oceani, sono fortemente alcalini. Dopo queste acquisizioni si ebbe un lungo intervallo che si protrasse fino agli ultimi anni cinquanta quando Lars Gunnar Sillén dell'Istituto reale di tecnologia di Stoccolma si accinse a chiarire la questione. Sillén si pose alcune domande: Che cosa controlla il pH, ossia la concentrazione di ioni idrogeno, degli oceani? Perché il pH è sempre compreso tra 7,5 e 8, ossia molto vicino alla neutralità acido-base? Successivamente Frederick T. MacKenzie e Robert M. Garrels della Northwestern University tradussero le domande abbastanza astratte di Sillén in termini geologici.

Consideriamo la disgregazione delle rocce vulcaniche causata sui continenti dagli agenti atmosferici. La pioggia che batte sulle rocce è una soluzione acida perché l'anidride carbonica dell'atmosfera si scioglie nell'acqua e forma acido carbonico (H_2CO_3). Da parte loro le rocce sono un reticolo di silicati di alluminio che incorporano cationi. L'acido reagisce con il reticolo in questo modo: cede ioni idrogeno trasformandosi in bicarbonato (HCO_3^-) e l'acqua piovana diviene così alcalina. Gli ioni idrogeno vanno a occupare nel reticolo il posto dei cationi, liberandoli in soluzione. Il reticolo viene quindi distrutto e si trasforma in minerali argillosi dalla struttura disordinata come la caolinite ($Al_2Si_2O_5(OH)_4$), facilmente soggetti all'erosione.

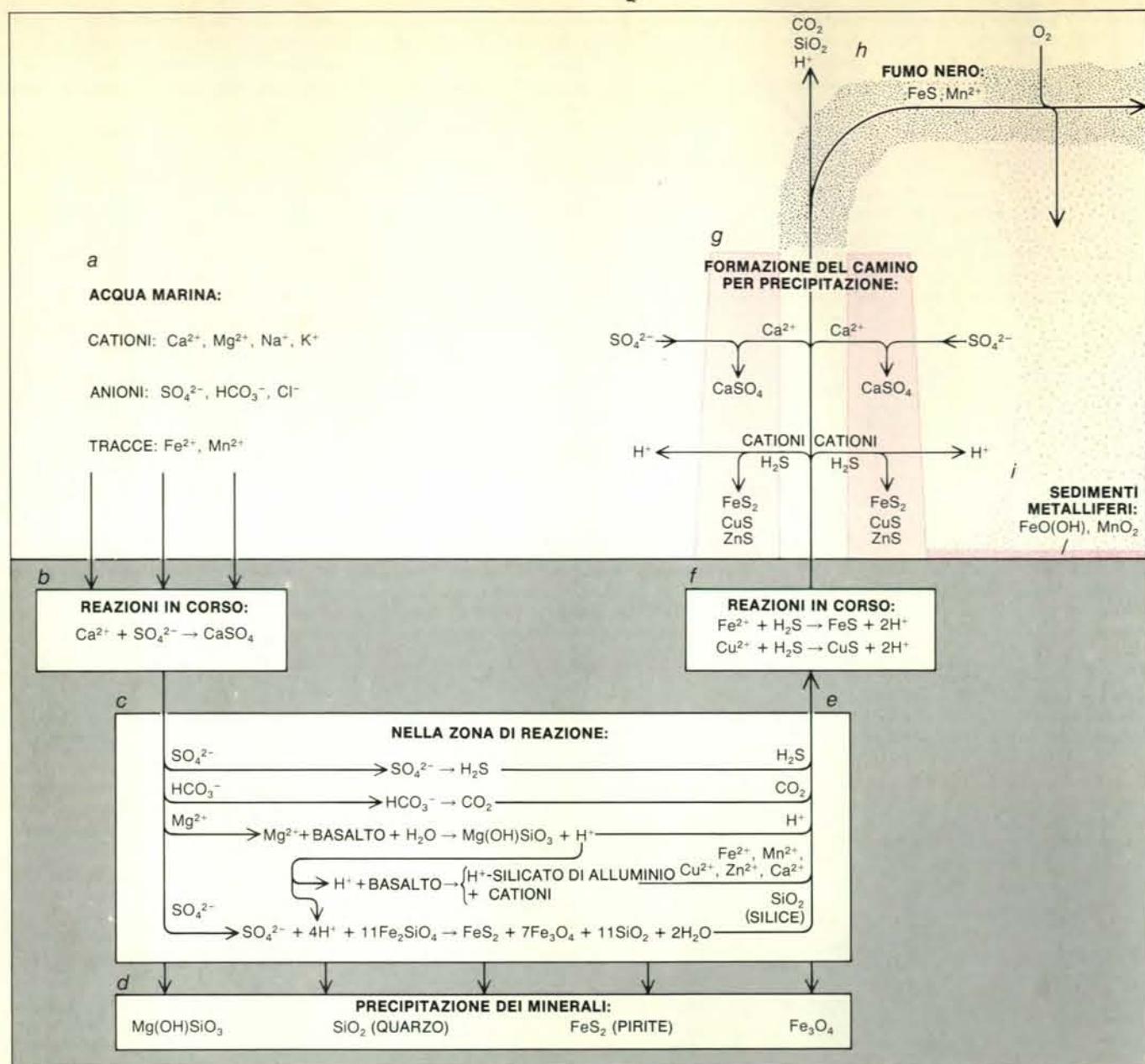
L'ipotesi di un equilibrio chimico tra atmosfera, continenti e oceani è stata avanzata negli anni cinquanta da Lars Gunnar Sillén dell'Istituto reale di tecnologia di Stoccolma e sviluppata nei particolari da Frederick T. Mackenzie e Robert M. Garrels della Northwestern University. Nell'atmosfera *(a)*, l'anidride carbonica si trova disciolta nel vapore acqueo come acido carbonico. L'acqua cade in forma di pioggia o neve sui continenti *(b)*, dove gli ioni idrogeno contenuti nella soluzione acida estraggono per lisciviazione dalle rocce continentali cationi (ossia ioni positivi) come calcio, magnesio, sodio e potassio. I fiumi trasportano i prodotti della disgregazione causata dagli agenti atmosferici negli oceani *(c)* nei quali, essendo poveri di ioni idrogeno, le reazioni si invertono dando origine a rocce sedimentarie ricche di cationi e restituendo all'atmosfera l'anidride carbonica.

Si può stimare la velocità globale di queste reazioni deducendola dal tasso a cui i loro prodotti sono trasportati dai fiumi negli oceani. Il risultato ottenuto esige che una quantità di anidride carbonica pari alla sua abbondanza nell'atmosfera debba essere consumata in circa 4000 anni. Evidentemente deve esserci una reazione che riconverte il bicarbonato in anidride carbonica. MacKenzie e Garrels notarono che il rapporto tra cationi e ioni idrogeno è di circa $1,2 \times 10^3$ nelle acque continentali medie, mentre è 6×10^7, ossia quattro ordini di grandezza in più, negli oceani. Quando, dunque, i silicati di alluminio ricchi di ioni idrogeno sono trasportati negli oceani (dove gli ioni idrogeno scarseggiano e i cationi abbondano) le reazioni che li hanno formati dovrebbero invertirsi. Dovrebbero formarsi argille ricche di cationi, gli ioni idrogeno essere liberati e poi combinarsi con ioni bicarbonato con la produzione di acqua e di anidride carbonica.

Le implicazioni erano profonde. La composizione grezza della crosta continentale si è mantenuta costante per almeno gli ultimi 2,5 miliardi di anni e, quindi, lungo lo stesso periodo, anche la composizione degli oceani non avrebbe dovuto variare. Furono intraprese adeguate ricerche per stabilire se effettivamente avessero luogo negli oceani le reazioni fondamentali che portano alla ricostituzione dei minerali argillosi. I risultati furono deludenti. Anche l'interpretazione più ottimistica dei dati non si avvicinava alla scala necessaria a raggiungere l'equilibrio tra la terraferma e il mare. A questo punto ci si trovò nel vicolo cieco ricordato prima. La via di uscita venne indicata dai calcoli preliminari condotti sui flussi chimici che potrebbero provenire dall'attività idrotermale presente lungo gli assi delle dorsali oceaniche.

Le prime esplorazioni

Naturalmente una cosa è sperare che le sorgenti calde siano un fenomeno comune nei centri di espansione oceanica e un'altra è trovarle effettivamente al di sotto di almeno 2500 metri d'acqua. Quando, all'inizio degli anni settanta, vennero fatti i primi tentativi, sia la conoscenza del fondo oceanico sia le tecniche di esplorazione erano rudimentali. I dispositivi sonar di quegli anni erano adeguati a rilevamenti su fondali piani, ma nella topografia accidentata tipica degli assi delle dorsali gli strumenti registravano un guazzabuglio incomprensibile di echi. Per questi motivi le prime apparecchiature calate a scopo esplorativo venivano in realtà trascinate lungo complicati percorsi a ostacoli nella più completa oscurità. Gli strumenti venivano danneggiati o perduti con una frequenza allarmante, ma, nonostante tutto, il dispositivo più sofisticato, il Deep Tow, gestito dalla Scripps Institution, fece importanti scoperte.

Il Deep Tow, un contenitore trainato da una nave di superficie per mezzo di un cavo telemetrico alla cui estremità era agganciato, trasportava telecamere, sonar, manometri e sonde per misurare la temperatura e la conducibilità elettrica dell'acqua (un parametro per valutarne la salinità). Per diversi anni il Deep Tow rilevò le anomalie termiche in numerosi punti al di sopra dei centri di espansione del Pacifico tropicale orientale. Nel corso di una esplorazione il Deep Tow trasportava anche campionatori d'acqua e Ray F. Weiss della Scripps fece in modo di prelevare un campione d'acqua dal centro di un'anomalia. La temperatura risultò inferiore di 0,1 gradi centigradi alla temperatura dell'acqua circostante (che era di due gradi centigradi). Eppure, i dati chimici, compresa la concentrazione di elio 3, dimostravano senza alcun dubbio che quell'acqua era di origine idrotermale. Presumibilmente il Deep Tow era capitato all'interno di un «pennacchio» che si tro-

Le reazioni idrotermali che avvengono nelle dorsali medio-oceaniche rigenerano anidride carbonica in quantità molto maggiore di quanto avverrebbe nel processo oceanico a bassa temperatura ipotizzato da Sillén. L'acqua marina *(a)* filtra attraverso la crosta oceanica e, forse, man mano che scende, parte del suo contenuto di ioni precipita *(b)* come solfato di calcio. Gli ioni restanti raggiungono una zona diversi chilometri al di sotto del fondo oceanico, dove, reagendo con la roccia basaltica molto calda *(c)*, provocano la precipitazione di minerali *(d)* con la formazione di una soluzione idrotermale calda, acida e ricca di metalli *(e)* che sale di nuovo verso il fondo oceanico. La soluzione nel corso della risalita può incontrare acqua marina fredda e raffreddarsi, dando così origine per precipitazione a solfuri metallici *(f)*. Poi, sul fondo oceanico, la soluzione si miscela con i solfati contenuti nell'acqua marina causando la precipitazione di solfuri metallici e solfato di calcio da cui sono formati i camini. Sopra i camini, il ferro, ancora in soluzione, forma il fumo che è nero per la presenza di solfuro di ferro *(h)*. Il manganese rimane in soluzione. Infine, il ferro e il manganese vengono ossidati e «piovono» sul fondo oceanico dando origine ai sedimenti metalliferi che si trovano ovunque al di sopra della crosta oceanica *(i)*. Nel frattempo l'anidride carbonica che fuoriesce dalle bocche termali si miscela con l'acqua dell'oceano, ne raggiunge in seguito la superficie, entra nell'atmosfera e chiude così il ciclo dell'anidride carbonica.

vava 15-20 metri al di sopra di una sorgente calda sottomarina.

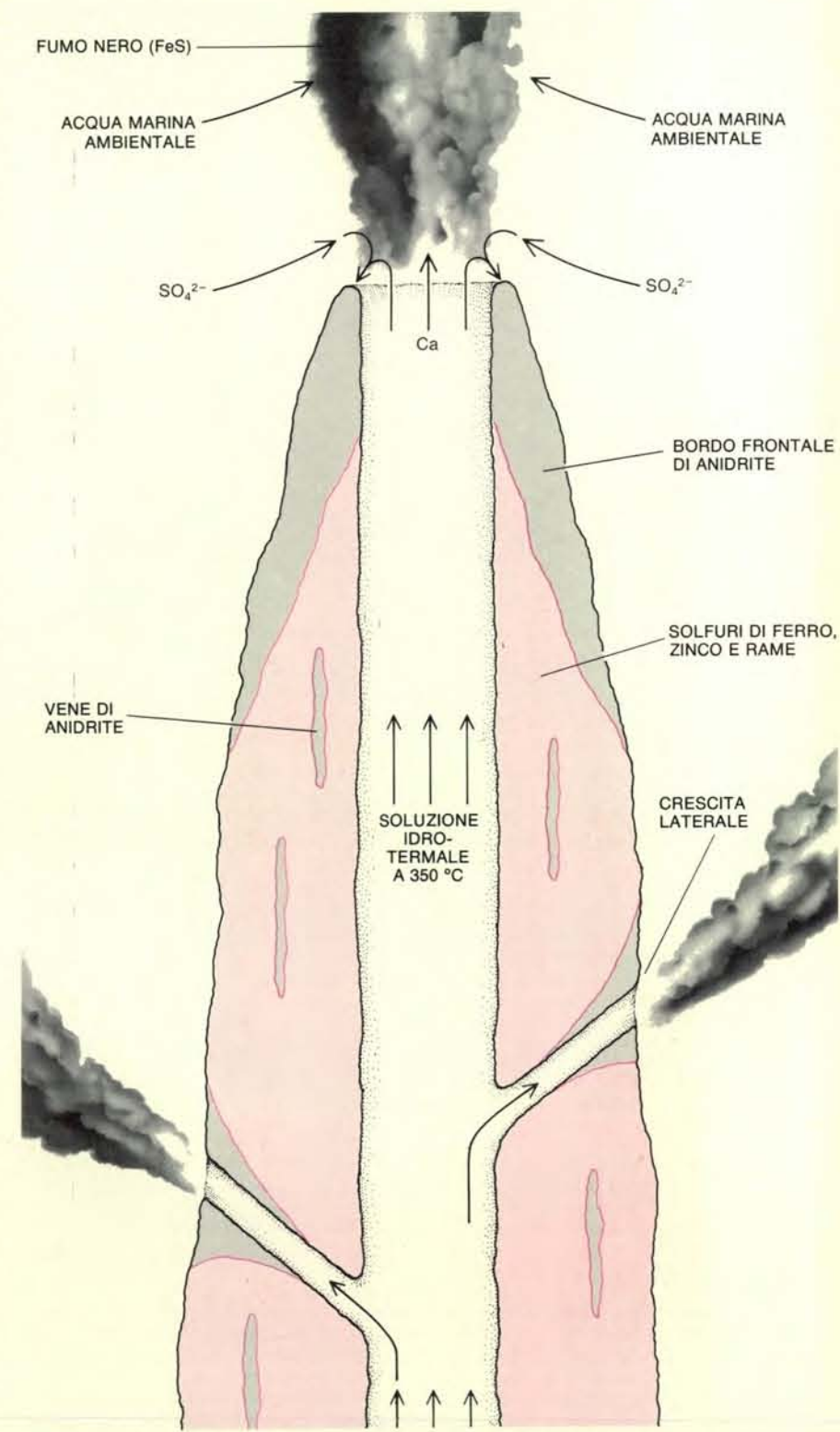
Verso la metà degli anni settanta le cose cominciarono rapidamente a cambiare per tre ordini di motivi. In primo luogo, l'US Navy mise a disposizione dei ricercatori le tecniche sviluppate per cartografare il fondo oceanico, e quindi entrarono in servizio di routine sistemi di navigazione d'alto mare di grande precisione. Questi sistemi si basano sulla misurazione del tempo che trascorre tra la trasmissione di un impulso acustico («domanda») da parte di un veicolo sottomarino e la ricezione di segnali acustici di ritorno («risposte») rimandati da una schiera di radiofari ormeggiati sul fondo. Se l'apparecchiatura sonar a bordo della nave di superficie riesce a determinare con esattezza le posizioni relative dei radiofari il natante sottomarino può essere pilotato con facilità. In realtà si possono ottenere le posizioni con un'accuratezza di qualche metro.

In secondo luogo, venne messa a punto una macchina da ripresa a campo largo. Gli apparecchi fotografici disponibili fino a quel momento erano leggeri e fragili e poiché venivano trainati sott'acqua a distanza dalla nave di superficie era molto difficile controllarli. Ciò che serviva era un «carrello» per le apparecchiature fotografiche pesante e si giunse così all'Angus, messo a punto dalla Woods Hole Oceanographic Institution. Si tratta di una enorme gabbia, pesante una tonnellata e mezzo, nella quale sono montati apparecchi fotografici a colori, lampade stroboscopiche, erogatori di potenza, sonar e un radiofaro per navigazione acustica. Nella sua configurazione attuale, il «carrello» viene trainato sott'acqua da una nave a una velocità di circa quattro chilometri all'ora, a un'altezza di circa 20 metri al di sopra del fondo marino. Dato il suo peso, non è mai a più di 75 metri dalla verticale della nave e, quindi, le letture del sonar di bordo possono contribuire a evitare che l'Angus si danneggi. Normalmente un'immersione dell'Angus dura 18 ore, durante le quali viene ripresa una fotografia a colori ogni 10 secondi. L'Angus viene poi sollevato, la pellicola sviluppata e segnalate le strutture morfologiche interessanti. Ogni fotografia reca la registrazione di quando è stata fatta; da questo dato la struttura può essere localizzata esattamente.

La terza tecnologia fondamentale che contribuì a modificare la situazione fu la ricerca sottomarina. I mezzi sommergibili che possono operare alle profondità degli assi delle dorsali sono di piccole dimensioni. Il più grande, l'*Alvin,* gestito dalla Woods Hole, pesa 16,5 tonnellate e può ospitare a bordo, nella sua sfera di titanio resistente alla pressione, del diametro di due metri, due ricercatori e un pilota. Ha una potenza molto limitata e, quindi, è lento; la sua velocità massima sul fondo è di circa quattro chilometri all'ora. Inoltre i fari di cui è dotato possono illuminare l'acqua entro una distanza di soli 15 metri circa. Si tratta dunque di un mezzo di esplorazione poco efficiente. Il suo impiego più appropriato è quello, però, di portarsi su luoghi già individuati come importanti ai fini dello studio. L'*Alvin* viene guidato sull'obiettivo per mezzo degli stessi radiofari di navigazione che servono a guidare l'Angus ed è possibile farlo arrivare sul fondo entro alcune decine di metri dal punto stabilito.

Uno scenario incredibile

Questo sistema elaborato e costoso che si basa su tre diverse tecnologie è stato impiegato per la prima volta nella primavera del 1977 in una dorsale di espansione del Pacifico, 280 chilometri a nord-est delle isole Galápagos. Giunse per prima sul posto la nave da ricerca *Knorr,* salpata da Woods Hole, e calò una rete di radiofari controllandone la posizione. Poi entrò in funzione l'Angus, sotto la direzione di Robert D. Ballard della Woods Hole. Prima che alcuni giorni dopo arrivasse l'*Alvin,* a bordo della piccola nave madre *Lulu,* erano già stati individuati diversi obiettivi. Nelle fotografie riprese dall'Angus comparivano in molti casi grandi bivalvi bianchi su un campo nero di basalto. L'*Alvin,* con a bordo due ricercatori, John B. Corliss dell'Oregon State University e uno di noi (Edmond) venne guidato verso uno di questi obiettivi.

L'*Alvin,* che affonda passivamente lungo la colonna d'acqua dell'oceano a una velocità di 30-35 metri al minuto, si fermò a una profondità di 2500 metri dopo un'ora e mezzo di discesa. A questo punto, a 100 metri dal fondo, l'*Alvin* si alleggerì sganciando una serie di zavorre e, «in folle», ci avvicinammo al fondo, un pendio dolce, dove per circa mezz'ora ci aggirammo cercando di individuare l'obiettivo. Ognuno di noi aveva a disposizione un oblò di Plexiglas attraverso cui guardare. Ci fermammo poi a raccogliere alcuni campioni di roccia e, mentre il pilota, servendosi del braccio meccanico del batiscafo, era alle prese con un cuscino di basalto (un deposito vulcanico che si forma per la lenta estrusione dal fondo oceanico di lava la cui superficie si raffredda, si indurisce, si frattura e si richiude gonfian-

Un fumatore nero si forma per accumulo di solfato di calcio che precipita da una soluzione idrotermale e forma un «bordo frontale» di anidrite biancastra. Man mano che il fronte avanza viene esposto alla soluzione calda non diluita. Man mano che si ridiscioglie, però, agisce da stampo per i solfuri metallici. L'anidrite si conserva nei posti in cui i solfuri la inglobano rapidamente.

dosi e assumendo così la forma di un cuscino) la nostra attenzione fu attirata da due enormi anemoni di mare color porpora. Solo quando spostammo lo sguardo, ci rendemmo effettivamente conto che l'acqua nel raggio dei fari tremolava come avviene per l'aria sopra un fondo stradale caldo. La temperatura dell'acqua misurata frettolosamente era superiore di cinque gradi centigradi a quella dell'ambiente (2,05 gradi centigradi). Dimenticandoci completamente delle rocce, prelevammo un campione d'acqua e poi continuammo a risalire il pendio. Fu così che arrivammo in vista di uno scenario incredibile.

La formazione basaltica tipica degli assi delle dorsali è proprio molto squallida: monotoni campi di cuscini scuri tagliati da faglie e da spaccature e per metri quadrati non si trova un solo organismo. Eppure, qui, vi era una vera e propria oasi. Scogliere di mitili e distese di bivalvi giganti ci apparvero nel tremolio dell'acqua insieme a granchi, anemoni di mare e ghiozzi. Le restanti cinque ore del tempo destinato alla ricerca furono frenetiche. Rilevammo la temperatura, la conducibilità, il pH e il contenuto di ossigeno dell'acqua; riprendemmo fotografie; prelevammo campioni d'acqua; ci assicurammo di aver prelevato esemplari di organismi sufficientemente rappresentativi, tutto nello stato d'ansia crescente dovuto alla diminuzione progressiva dell'energia a disposizione del nostro equipaggiamento.

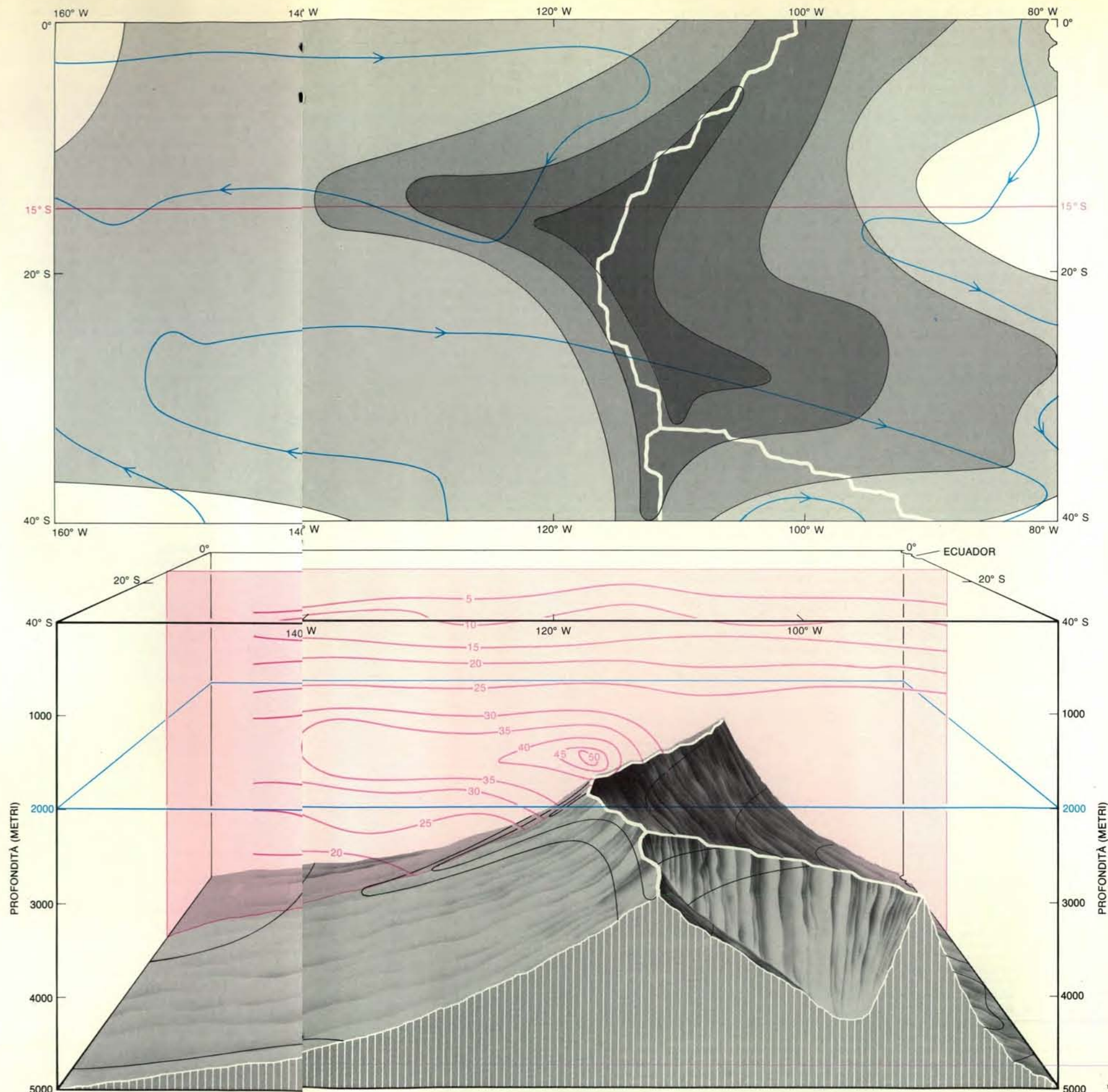
Per fortuna tutto funzionò senza inconvenienti. Fu subito evidente che eravamo capitati in un campo di sorgenti calde. Su un'area circolare di circa 100 metri di diametro, l'acqua calda sgorgava da ogni orifizio e da ogni spaccatura del fondo oceanico. La temperatura dell'acqua variava molto, ma il massimo era di circa 17 gradi centigradi. Gli organismi erano selettivi al massimo; ingombravano infatti i condotti idrotermali più caldi e addirittura le scogliere di mitili circondavano il flusso dell'acqua formando esse stesse dei condotti.

Lavorammo finché l'energia disponibile per le apparecchiature scientifiche si esaurì, scaricammo una seconda serie di zavorre e l'*Alvin* lasciò il fondo. L'equipaggiamento venne fatto rientrare e presto cominciammo a sentire freddo. Dopo otto ore durante le quali eravamo stati rannicchiati nei due metri della sfera in condizioni in cui ogni movimento aveva richiesto la collaborazione dei colleghi altrettanto intralciati, trascorremmo i novanta minuti di risalita cercando di scaldarci, troppo stanchi perfino per osservare gli organismi bioluminescenti disturbati dal passaggio del batiscafo. Come sempre, l'avvicinamento alla superficie fu un ritorno alla vita. A circa 200 metri di profondità, l'acqua venne pervasa da una debole luminescenza verde e poi, subito dopo, tutto divenne luminoso. Il pilota aprì le casse di zavorra e l'*Alvin* fu così circondato da una colonna di bolle. Vi fu un urto soffice quando arrivammo in superficie e poi un pigro ondeggiare. In breve i sommozzatori circondarono il batiscafo per attaccarvi i cavi di recupero con cui l'*Alvin* venne riportato nel suo alloggiamento. Toccò a noi a questo punto alzarci e uscir fuori. Dal momento in cui barcollanti entrammo a bordo della *Lulu*, tutti condivisero la nostra eccitazione. Il cesto dell'*Alvin* era carico di bivalvi giganti e di mitili. I granchi, penetrati sotto il rivestimento in fibra di vetro del batiscafo, cominciarono a cadere sul ponte.

Vita basata sull'acido solfidrico

A questo punto cominciò la ricerca vera e propria. I campionatori vennero smontati e i campioni d'acqua, trasferiti sulla *Knorr*, vennero analizzati durante la notte. Le fotografie riprese dall'Angus vennero esaminate per individuare altri obiettivi; i campionatori vennero puliti e rimontati, cosicché alle ore 6 del giorno successivo l'*Alvin* era già pronto. Alle 8 e 45 si immerse di nuovo. In totale compì 15 immersioni.

Man mano che il lavoro di analisi proseguiva, divenne chiaro che tutte le acque termali avevano un'elevata concentrazione di acido solfidrico, o solfuro d'idrogeno (H_2S), e ciò spiegava le oasi trovate sul fondo oceanico. Batteri che traggono energia dall'ossidazione dell'acido solfidrico sono comuni in molti sistemi ecologici; nelle oasi, però, essi dovevano essere i produttori primari alla base di una piramide ecologica. Fondamentalmente, l'energia che alimenta questo nuovo ecosistema non proviene dal Sole, come nella fotosintesi, ma dal decadimento radioatti-

La distribuzione dei sedimenti metalliferi sul fondo del Pacifico meridionale, così come è stata cartografata da Kurt G. T. Boström della Scripps Institution of Oceanography, presenta configurazioni che la rendono asimmetrica rispetto all'asse della dorsale da cui i metalli scaturiscono in soluzione idrotermale. Il disegno in alto mostra il Pacifico meridionale visto dall'alto. A 15 gradi di latitudine sud un saliente lungo e stretto (rappresentato con varie tonalità di grigio) di sedimenti ricchi di ossidi di ferro e manganese si proietta verso ovest rispetto all'asse. A 30 gradi di latitudine sud, si protende verso est un saliente più largo e più corto. Le curve in blu indicano le correnti oceaniche a una profondità di 2000 metri (600 metri circa sopra l'asse della dorsale) come sono state desunte da Joseph L. Reid della Scripps Institution. Le correnti hanno allontanato gli ossidi dall'asse, e quindi spiegano la formazione dei salienti. Il disegno in basso rappresenta il Pacifico meridionale visto da sud. Le curve in rosso segnalano le concentrazioni dell'isotopo elio 3 in un pennacchio che si estende verso ovest attraverso il Pacifico a 15 gradi di latitudine sud. L'andamento dei valori è stato misurato da John E. Lupton e Harmon Craig, anch'essi della Scripps. I numeri indicano la concentrazione di elio 3 come percentuale in eccedenza rispetto al normale valore marino di saturazione. Evidentemente l'elio è liberato dalla crosta oceanica attraverso reazioni idrotermali e, dato che è un elemento inerte, non subisce alcun cambiamento mentre sgorga lungo l'asse e caratterizza l'acqua oceanica da cui precipitano i sedimenti metalliferi.

vo di isotopi a lunga vita dell'uranio, del torio e del potassio che si trovano nell'interno della Terra. Il decadimento libera calore, il quale produce magma, che, a sua volta, diventa nuova crosta oceanica. L'acqua marina filtra dentro la crosta e reagisce con le rocce a temperatura e a pressione elevate.

Per le oasi sono fondamentali due reazioni. I solfati contenuti nell'acqua marina reagiscono con il ferro presente nelle rocce e producono acido solfidrico e ossidi di ferro; inoltre i solfuri delle rocce vengono disciolti. Per questo motivo l'acqua calda che sale ritornando verso il fondo oceanico è molto ricca di acido solfidrico. Sul fondo, nei pressi delle bocche termali, l'acido solfidrico viene assorbito dai batteri i quali assumono anche l'ossigeno di origine fotosintetica presente nell'acqua marina. Ione solfuro e ossigeno si combinano, cosicché il solfuro si trasforma in solfato. La reazione libera l'energia che alimenta il metabolismo dei batteri, i quali, a loro volta nutrono altre specie. La varietà delle specie fu un'indicazione diretta che le sorgenti calde devono essere comuni negli assi delle dorsali in tutto il mondo. In quale altro posto potrebbe svilupparsi e mantenersi una fauna così unica ed evoluta?

Al nostro laboratorio del Massachusetts Institute of Technology venne affidato il compito di analizzare in dettaglio i campioni d'acqua che l'*Alvin* aveva prelevato alle bocche termali. Chiaramente si vide che la chimica dell'acqua era il risultato di una complessa serie di reazioni avvenute tra l'acqua marina e il basalto di una zona di reazione inaccessibile, parecchie migliaia di metri al di sotto del fondo oceanico. Di conseguenza il modo migliore per procedere nel «dipanare» le reazioni sarebbe stato quello di determinare in che grado i campioni d'acqua erano ricchi o poveri dei vari elementi chimici rispetto al basalto e rispetto all'acqua marina comune. Alla fine avevamo misurato la concentrazione di 35 elementi, probabilmente la maggior quantità mai misurata in una ricerca di questo tipo. Man mano che si procedeva, la dinamica chimica cominciò a chiarirsi.

Le proprietà delle acque termali

Magnesio e solfato, che sono costituenti dell'acqua marina comune, mostrarono una diminuzione uniforme di concentrazione con l'aumentare della temperatura (ossia, la temperatura di ciascun campione d'acqua così come era stato prelevato dall'*Alvin*). Ovviamente ogni campione era costituito di acqua che era stata emessa dai condotti termali e quindi diluita in una certa misura dall'acqua marina dell'ambiente. L'estrapolazione a una concentrazione zero delle tendenze concentrazione-temperatura, basata sull'ipotesi che acqua esclusivamente termale dimostrasse di non contenere né magnesio né solfato, portò a una temperatura di circa 350 gradi centigradi. L'estrapolazione era dichiaratamente stravagante: l'acqua più calda che abbiamo prelevato aveva una temperatura di soli 19 gradi. D'altra parte, la concentrazione di silice (quarzo disciolto in acqua) aumentava all'aumentare della temperatura. L'estrapolazione della tendenza a questo punto di intersezione con le curve relative alla solubilità del quarzo in acqua a varie pressioni e temperature portò anch'essa a una temperatura di circa 350 gradi centigradi.

Anche il ferro, il rame, il nichel e il cadmio, che si combinano con l'acido solfidrico presente nell'acqua termale e formano solfuri od ossidi che precipitano dall'acqua, mostrarono diminuzioni marcate con l'aumentare della temperatura, e scesero a zero in un intervallo di temperature compreso tra 30 e 35 gradi centigradi. Questo intervallo poteva essere interpretato come la temperatura di un serbatoio d'acqua spillato dalle sorgenti calde. Supponemmo che l'acqua che sale da un regime di reazioni ad alta temperatura trascinasse con sé «acqua sotterranea» la cui temperatura e composizione fossero simili a quelle dell'acqua marina ambientale che si trova sopra l'asse delle dorsali. La soluzione ad alta temperatura, ossia lo stadio «finale» del sistema, verrebbe raffreddata da questa acqua sotterranea e la sua acidità diluita, così che si avrebbe la precipitazione degli elementi che formano i solfuri e gli ossidi dell'«impianto idraulico» della crosta basaltica. La precipitazione procederebbe fino a un tale grado che gli elementi che formano i solfuri e gli ossidi verrebbero rimossi perfino dall'acqua sotterranea. Sarebbero stati forniti, dunque, ai nostri campioni soltanto dall'acqua ambientale che li aveva diluiti. Le prove delle reazioni ad alta temperatura erano presenti però nei dati relativi al magnesio, al solfato e alla silice.

L'elio 3 aumentava molto con la temperatura: $2,2 \times 10^{-17}$ moli (circa 13 milioni di atomi) per caloria. Nell'elio rimasto, una volta sottratta la quantità di equilibrio dalla concentrazione misurata, il rapporto tra elio 3 ed elio 4 era otto volte quello relativo all'atmosfera. Sulla scia della scoperta della differenza riguardante l'elio 3 nelle profondità del Pacifico, Craig, Clarke e i loro colleghi avevano calcolato che la quantità di elio 3 necessaria perché la differenza esistesse a livello mondiale compensando la perdita di elio verso l'atmosfera durante il rimescolamento degli oceani era di circa 1100 moli per anno. Se tutto l'elio 3 che si immette negli oceani passa attraverso le sorgenti calde degli assi delle dorsali e se il rapporto tra elio 3 e calore è dovunque lo stesso misurato nell'asse della dorsale delle Galápagos, il trasporto mondiale di calore nelle sorgenti calde degli assi delle dorsali è facilmente calcolabile. Ammonta a 5×10^{19} calorie per anno e questo valore si colloca al centro dell'intervallo di valori che i geofisici hanno calcolato.

Potrebbe essere giustificato, quindi, supporre che le anomalie riscontrate nei nostri campioni a proposito della concentrazione di altri elementi servissero (rapportate all'anomalia dell'elio 3) a calcolare i flussi totali di quegli elementi immessi dalle sorgenti calde sottomarine negli oceani. I risultati sono impressionanti, soprattutto se vengono messi a confronto con i flussi provenienti dalla disgregazione causata dagli agenti atmosferici sui continenti. Gli assi delle dorsali consumano quindi la maggior parte del magnesio e dei solfati che i fiumi immettono negli oceani, mentre liberano da cinque a 10 volte più litio e rubidio e tra un terzo e la metà di potassio, calcio, bario e silice. La quantità di magnesio liberata è sufficiente a spiegare tutto l'accumulo di quell'elemento nei sedimenti metalliferi e nei noduli che tappezzano l'oceano. Infine, gli assi delle dorsali ritrasformano in anidride carbonica la maggior parte del bicarbonato prodotto dalla disgregazione causata dagli agenti atmosferici sui continenti. Le reazioni oceaniche a bassa temperatura che Sillén aveva previsto vengono quindi sostituite da un processo idrotermale ad alta temperatura.

Le ofioliti, segmenti di crosta oceanica che si trovano sopra i continenti, includono strati di sedimenti identici a quelli che tappezzano il fondo oceanico. Contengono anche giacimenti minerari di solfuri che si suppone siano stati prodotti sul fondo oceanico per attività idrotermale di molti fumatori neri. Spesso si trovano in depressioni sopra uno strato di basalto.

I «fumatori neri»

Dalle analisi condotte apparve chiaro che andava modificato ciò che si pensava dei processi che alla scala dei tempi geologici determinano la chimica dell'acqua marina. Naturalmente sarebbe stato opportuno scoprire le soluzioni calde - ossia lo «stadio finale» del sistema - così come fuoriescono dal fondo oceanico in forma non diluita. La prospettiva, però, era resa molto incerta dall'estesa presenza di acqua sotterranea nell'asse della dorsale delle Galápagos. Ci sorprese, dunque, e ci emozionò quello che udimmo da Jean Francheteau dell'Institut de Physique du Globe. Il batiscafo da ricerca francese, *Cyana,* aveva trovato estesi depositi di solfuri nella cresta della Dorsale pacifica orientale subito a sud dell'imboccatura del Golfo di California a 21 gradi di latitudine nord. Nel luogo non si erano trovate tracce di attività idrotermale attuale. Ciononostante i depositi avrebbero potuto formarsi solo per precipitazione di solfuri da soluzioni molto calde.

Noi stessi ritornammo all'asse della dorsale delle Galápagos nella primavera del 1979, dove trovammo altri campi termali, ma la temperatura non superava mai i 23 gradi centigradi. L'*Alvin* allora venne trasferito nei pressi del Golfo di California. L'Angus, già sul posto, aveva fotografato campi di sorgenti calde e in mezzo a essi l'*Alvin* trovò quello che non pensavamo mai di trovare: grandi getti di acqua nera sgorgavano da camini di solfuri che si erano autoformati e raggiungevano un'altezza di parecchi metri. La temperatura dell'acqua termale superava i 300 gradi centigradi. (Alla pressione esistente al livello del mare l'acqua a quella temperatura bollirebbe in maniera esplosiva, come avviene nei geyser tipo l'Old Faithful di Yellowstone, mentre alla pressione esistente a quelle profondità l'acqua si mantiene liquida.) Sorse il problema di progettare e costruire campionatori d'acqua che potessero funzionare a temperature superiori ai 300 gradi centigradi; la National Science Foundation destinò altro tempo alle immersioni dell'*Alvin* dirette ai cosiddetti «*black smoker*» (fumatori neri); John A. Archuleta del Los Alamos National Laboratory ci prestò l'equipaggiamento.

Ci immergemmo nel novembre 1979. I radiofari di navigazione lasciati sul posto ci guidarono verso i condotti termali e così ci avvicinammo a un fumatore nero che si innalzava sul fondo oceanico in mezzo a grandi blocchi di solfuri. Manovrando in una forte corrente di fondo, il pilota inserì una sonda di nuova progettazione per misurare la temperatura proprio nella bocca del fumatore, che aveva un diametro di soli 15 centimetri. La lettura si stabilizzò entro alcuni decimi di grado sui 350 gradi centigradi. La capacità previsionale della chimica era riaffermata! Tutti i condotti termali che campionammo avevano una temperatura che non si allontanava dai 350 gradi se non di qualche unità.

Le soluzioni a 350 gradi centigradi escono dai condotti sotto forma di fluidi limpidi e omogenei. Contengono 100 parti per milione di ferro e alcune parti per milione di zinco, rame e nichel. Ciò significa arricchimenti per fattori dell'ordine di 10^8, mentre la concentrazione di ferro, zinco, rame e nichel nell'acqua marina comune si misura in parti per bilione. Inoltre le soluzioni contengono 210 parti per milione di acido solfidrico, mentre l'acqua comune dell'oceano non ne contiene. La concentrazione di silice è di 1290 parti per milione, un valore esattamente conforme a quello prevedibile dopo l'esame dei dati relativi alla dorsale delle Galápagos. In altre parole, le soluzioni sono sature di quarzo, mentre non contengono né magnesio, né solfato, in accordo ancora una volta con le previsioni.

Quando giungono alla bocca del condotto, le soluzioni si mescolano velocemente con la fredda acqua marina dell'ambiente, dando origine così al «fumo» nero. Si tratta di una sospensione di fini particelle di solfuro di ferro che precipitano dalle soluzioni. Sembra che i camini si formino per la crescita di un bordo frontale di solfato di calcio. Il calcio viene fornito dalle soluzioni idrotermali e il solfato dall'acqua marina. Man mano che il bordo avanza si espone alle acque termali non diluite e, quindi, si scioglie di nuovo e viene rimpiazzato da solfuri precipitati. L'acqua termale comunemente si apre un varco nelle fenditure del camino favorendone la crescita laterale. Si ha così un deposito costituito da diversi minerali.

Lo sfruttamento dei minerali metalliferi

Non varrebbe la pena sfruttare direttamente le soluzioni, Il recupero di una tonnellata di zinco, per esempio, richiederebbe il trattamento di circa sei milioni di tonnellate di soluzione, la quantità che fluisce da un fumatore nero tipico in un periodo di parecchi mesi. Una tonnellata di zinco vale circa 900 dollari, ed è quindi meglio che la natura segua il suo corso e sfruttare i depositi già formatisi.

Consideriamo in primo luogo i fumatori neri. La formazione di un massiccio deposito di solfuri, come quelli delle ofioliti, che contiene milioni di tonnellate di minerali, sembrerebbe richiedere una «foresta» di fumatori. Forse i fumatori occupano inizialmente una depressione del fondo oceanico che poi verrebbe colmata dalle particelle che precipitano dal fumo. In ogni caso, il deposito idrotermale che si accumula attorno ai fumatori sarà sottoposto a successivi trattamenti da parte delle soluzioni calde che salgono attraverso il deposito verso il fondo oceanico. Il deposito sarà privato degli elementi secondari (rame, zinco, cobalto e nichel) e rimarrà una matrice di grandi cristalli di pirite pura. Gli elementi secondari verranno depositati di nuovo sopra la pirite in forma di camini sul fondo oceanico. Nel frattempo, l'infiltrazione di acqua fredda nel deposito che si sta accumulando favorirà la precipitazione *in situ* di altri depositi dalle soluzioni idrotermali che stanno salendo.

Questa sequenza spiega le caratteristiche generali dei depositi che sono stati sfruttati nelle formazioni ofiolitiche. Il corpo principale di un deposito del genere è composto di pirite pura a grana grossolana. Lo strato che si trova sopra la pirite, che ha ossidato l'ocra, è ricco di elementi secondari. Inoltre, Alexander Malahoff della National Oceanographic and Atmospheric Administration, ha recentemente scoperto una vera e propria foresta di fumatori neri. A circa 30 chilometri dai campi delle Galápagos osservò dall'*Alvin* una moltitudine di camini estinti distribuiti per 600 metri lungo l'asse della dor-

Si pensa che i depositi tipo Besshi, giacimenti minerari inglobati in una roccia sedimentaria, si formino sotto l'oceano quando l'asse di una dorsale vicino a una massa continentale viene sepolto da detriti erosionali. I solfuri ricchi di metalli si accumulano centinaia di metri al di sopra dei dicchi e dei filoni basaltici, ossia intrusioni verticali e orizzontali di magma le quali danno origine ad attività idrotermale man mano che, dopo essere penetrate nei sedimenti, si raffreddano.

sale. Si trovavano sulla sommità di un giacimento di solfuri alto parecchie decine di metri. Considerata l'enorme quantità di calore che i camini devono aver liberato quando erano in attività, probabilmente a quel tempo un batiscafo da ricerca non avrebbe potuto avvicinarli. Basti pensare che gli oblò di Plexiglas dell'*Alvin* fondono a 86 gradi centigradi.

Va detto che la formazione di giacimenti minerari nelle località stesse dei fumatori neri è molto aleatoria. Una schiacciante frazione dei metalli disciolti nelle soluzioni idrotermali diventano particelle del fumo che esce dai camini; questa parte viene quindi dissipata dalle correnti oceaniche di profondità e, alla fine, le particelle reagiscono con l'ossigeno disciolto nell'acqua marina. Gli ossidi di ferro e di manganese non sono solubili in acqua e, di conseguenza, formano sedimenti metalliferi, i quali includono una certa quantità di rame, zinco, cobalto e nichel metallici, ma ciononostante sono troppo poveri per essere sfruttati commercialmente.

Dove si depositano esattamente i sedimenti ricchi di metalli? Sulla scia della scoperta fatta (insieme a Peterson), che i sedimenti metalliferi si trovano ovunque sul fondo oceanico, Boström si accinse a cartografare la distribuzione mondiale del contenuto in metalli dei sedimenti. Sul fondo del Pacifico meridionale trovò andamenti peculiari. A 15 gradi di latitudine sud un saliente lungo e stretto di sedimenti relativamente ricco di metalli si estendeva verso ovest rispetto all'asse della Dorsale pacifica meridionale. A 30 gradi di latitudine sud un saliente più ampio piegava verso est. Successivamente, John E. Lupton e Harmon Craig della Scripps Institution cartografarono uno spettacolare pennacchio di elio 3 che si estendeva verso ovest rispetto all'asse della dorsale attraversando quasi tutto il Pacifico a 15 gradi di latitudine sud.

Queste diverse osservazioni non ebbero spiegazione fino a una pubblicazione recente di Joseph L. Reid della Scripps. Egli aveva compiuto un rilevamento della densità dell'acqua in vari punti degli oceani e dai valori di densità aveva dedotto le forze di taglio esistenti tra gli strati oceanici. Poi, supponendo che l'oceano abbia «superfici» dove l'acqua si mantiene ferma (queste superfici sono i confini tra correnti costanti che fluiscono in direzioni diverse) è stato in grado di calcolare la grandezza e la direzione di flusso rispetto a una data superficie. Reid ha così cartografato il movimento dell'acqua nel Pacifico meridionale a una profondità di 2000 metri rispetto a un «livello immobile» che si è supposto giaccia a 3500 metri di profondità. Straordinariamente le asimmetrie rilevate da Boström coincidono esattamente con le linee di flusso dedotte da Reid. Anche il pennacchio di elio 3 viene spiegato. L'elio è un elemento chimico inerte e quindi mentre si sposta con le correnti oceaniche non subisce alcuna trasformazione. In effetti caratterizza l'acqua oceanica dalla quale l'effluente idrotermale si deposita sul fondo.

I depositi tipo Besshi

Se le scoperte fatte negli assi delle dorsali contribuiscono molto alla comprensione dei processi che danno origine ai giacimenti minerari ofiolitici, contribuiscono altrettanto alla comprensione di un tipo di deposito continentale di importanza economica maggiore. Si tratta dei «depositi tipo Besshi», chiamati così dal nome del deposito trovato in Giappone. In un deposito tipo Besshi, un massiccio giacimento di solfuri giace in mezzo a sedimenti, generalmente argillite, una roccia che ha origine da argilla a grana fine. Il giacimento minerario non si trova sopra il basalto. Infatti, spesso si deve trivellare l'argillite fino a profondità di parecchie centinaia di metri al di sotto del giacimento prima di trovare il basalto e anche allora il basalto è sotto forma di dicchi e di filoni, ossia intrusioni verticali e orizzontali di magma.

Evidentemente un deposito tipo Besshi si forma soltanto quando l'asse di una dorsale è vicino a una massa continentale che è fonte di una grande quantità di detriti erosionali. Vi è un caso in cui il processo è in corso. Nel Guaymas Basin, al centro del Golfo di California, la Dorsale pacifica orientale penetra nella zolla continentale e i sedimenti fluviali provenienti dal Messico hanno sepolto l'asse nel silt fino a una profondità di parecchie centinaia di metri. In queste condizioni, la messa in luogo di crosta oceanica di nuova formazione segue una via del tutto particolare. Il magma in ascesa non sbocca nel fondo oceanico, ma sale entro il silt, formando così dicchi e filoni. Il silt è permeabile all'acqua e, quindi, l'attività idrotermale è molto estesa e, infatti, il sito fu individuato, da Lawrence A. Lawver della Scripps, a causa del grande flusso sottomarino di calore che egli aveva misurato in quell'area.

Successivamente Peter F. Lonsdale della Scripps trovò, al centro del Guaymas Basin, cumuli di solfuri, ognuno dei quali alto parecchie decine di metri e lungo centinaia. La fusione delle camicie in plastica delle carote prelevate da Lonsdale fece pensare che a soli 10 metri al di sotto del fondo oceanico vi fosse una temperatura di 100 gradi centigradi. John Lupton della Scripps rilevò elevate concentrazioni di elio nelle acque oceaniche sovrastanti il sito. Perforazioni profonde attraversarono filoni basaltici molto caldi.

Nel gennaio dello scorso anno prendemmo parte a una spedizione condotta da Lonsdale nell'area. Nel corso di sette immersioni l'*Alvin* prelevò campioni di acqua che sgorgava da tre grandi cumuli. Le acque più calde avevano una temperatura di 315 gradi centigradi; la bassa concentrazione di magnesio e di solfato riscontrata indicava che si trattava dello stadio finale del sistema. In generale però la composizione rendeva quest'acqua estremamente diversa dagli stadi finali emessi dagli assi delle dorsali in oceano aperto. Senza dubbio è la loro risalita attraverso il silt del Guaymas Basin la causa della differenza. In particolare i gusci di carbonato di calcio degli organismi planctonici sepolti nel silt sono disciolti dallo stadio finale del Guaymas Basin, il quale, quando lascia la zona di reazione basaltica, deve essere molto acido (come nel caso degli stadi che fuoriescono dai fumatori neri a 21 gradi di latitudine nord). L'aggiunta di carbonato, che rende la soluzione alcalina, provoca la precipitazione dei solfuri. Quando lo stadio finale del Guaymas Basin entra nell'oceano è circa quattro volte più alcalino dell'acqua marina ed è quasi privo di ferro e di altri elementi che possano dare origine a giacimenti minerari.

Odore di gasolio

Nel Guaymas Basin, dunque, la differenza fondamentale è realmente costituita dalla ricchezza di organismi planctonici che caratterizza le acque oceaniche sovrastanti. Nel Guaymas Basin lo stadio finale del sistema incontra sotto il fondo oceanico un ambiente alcalino e quindi il contenuto metallico della soluzione è intrappolato molto efficacemente. In una dorsale in pieno oceano, come l'asse che si trova a 21 gradi di latitudine nord, lo stadio finale trova un ambiente alcalino solo quando incontra l'acqua marina. Si ha così la produzione di fumo nero e la dispersione di gran parte del contenuto metallico.

Nel Guaymas Basin, sembra che la ricchezza di plancton abbia provocato una curiosa modificazione dell'ecosistema idrotermale. Quando Lonsdale prelevò carote di sedimenti in quel bacino, notò che odoravano fortemente di gasolio. Quando vi ritornò lo scorso anno, i campionatori dell'*Alvin* furono ostruiti più volte da gocce di cera. La spiegazione fu facile. Il silt del Guaymas Basin è ricco di carbonio planctonico, il quale viene «scisso» in idrocarburi dal calore dei campi idrotermali. I campi del Guaymas Basin sono caratterizzati da grandi «tappeti» di batteri, i quali probabilmente si nutrono di idrocarburi. Sembra che i batteri attorno ai condotti termali competano con i vermi tubicoli. Mancano bivalvi, pesci e piccoli granchi; mentre invece sono comuni grandi granchi del genere *Stenocionops* che sembra predino i vermi. I campi termali del Guaymas Basin sono diversi sotto tutti gli aspetti dai campi degli assi delle dorsali che si trovano in oceano aperto e questo fa prevedere per le future scoperte una grande varietà di manifestazioni.

Gli ultimi due decenni sono stati testimoni di una vera e propria esplosione nelle scienze della Terra e nelle scienze planetarie. Per quanto riguarda la Terra, la teoria della tettonica a zolle ha unificato le scienze geologiche fino ad allora frammentarie e descrittive. Inoltre la scoperta dell'attività idrotermale lungo gli assi delle dorsali sottomarine dimostra che la novità e l'unicità non sono dominio esclusivo delle missioni spaziali dirette ad altri pianeti. Si dice che la Terra sia il pianeta dell'acqua. Ora si può dire che la chimica dell'acqua e dei sedimenti oceanici porta soprattutto il segno dei processi vulcanici.

I primi agricoltori della pianura baltica

Scavi compiuti in Polonia hanno portato in luce i resti di uomini dediti 7000 anni fa alle coltivazioni, agli allevamenti e alla caccia. Fu una fase di transizione, dopo la quale l'agricoltura si affermò stabilmente

di Peter Bogucki e Ryszard Grygiel

Uno degli eventi principali nel corso della preistoria umana - la nascita dell'agricoltura e dell'allevamento animale - si verificò in tempi diversi in varie parti del Vecchio Mondo. Nell'Europa meridionale la conoscenza di questo nuovo modo di vita potrebbe essersi diffusa in modo graduale a partire da quelle regioni del Vicino Oriente in cui si era già affermata la Rivoluzione neolitica. A nord dei Carpazi e delle Alpi, la diffusione del Neolitico fu invece un movimento di persone più che di idee e una delle poche migrazioni del genere ben documentate nei reperti archeologici.

Ciò nonostante, anche dopo generazioni e generazioni di lavoro sul campo, non sappiamo molto sulle persone che, partendo dalla valle del medio corso del Danubio, si addentrarono nel bassopiano ricoperto di foreste dell'Europa centro-settentrionale portando con sé le prime piante coltivate e un certo numero di animali addomesticati. La scarsità delle informazioni di cui disponiamo è dovuta in parte al fatto che la natura acida dei suoli di queste regioni portò alla rapida distruzione dei resti tanto umani quanto animali e in parte al fatto che gli archeologi hanno preferito lavorare su siti neolitici più estesi e di epoca posteriore anziché su siti più circoscritti e più antichi. Nondimeno, negli anni trenta alcuni siti del Neolitico inferiore, nella pianura polacca a circa 150 chilometri a nordovest di Varsavia, avevano cominciato a fornire informazioni preziose, ma le ricerche furono interrotte dallo scoppio della seconda guerra mondiale. Gli stessi siti, fatti oggetto recentemente di altre ricerche, continuano a fornire informazioni ancora oggi. Prima di descrivere i siti e i loro contenuti, dovremmo però dare al lettore un'idea della natura fisica di quest'ampia fascia al di là dei Carpazi e delle Alpi e delle opportunità che essa offrì ai primi «agricoltori» che vi migrarono dal sud.

Considerata a grandi linee, l'Europa centro-settentrionale può essere suddivisa in tre zone ecologiche principali. Quella più a sud, formata dalle aree montane dei Carpazi e delle Alpi, non fu in generale colonizzata da comunità di agricoltori sino al Neolitico avanzato, e pertanto non verrà discussa qui. Più a nord, oltre le montagne, si estende una zona di altipiani ondulati che vanno dalla Slovacchia e dalla Polonia meridionale, passando per la Germania centrale, sino all'interno della Francia. La regione è ricoperta in gran parte da löss, lo strato fertile, ma arido, di suolo polveroso depositato dai venti che spirarono attraverso l'Europa verso la fine dell'Epoca glaciale. Il löss riempie i bacini fluviali di questa regione e ricopre le colline interposte, raggiungendo in alcuni luoghi uno spessore di 100 metri.

La terza zona, quella più settentrionale, è formata dai bassipiani della pianura nordeuropea, e si estende dall'Unione Sovietica, verso ovest, alla Polonia, alla Germania settentrionale e alla Danimarca per arrivare sino ai Paesi Bassi. Gli strati di ghiaccio dell'ultima Epoca glaciale la ricoprirono sino a circa 15 000 anni fa; la loro avanzata scavò laghi lunghi e profondi e il loro ritiro lasciò un terreno ricoperto da paludi, depressioni riempite da laghi e un suolo ricoperto da sabbie, argille e ghiaie. Alcuni di questi nuovi suoli sarebbero diventati molto fertili; altri avrebbero avuto una fertilità solo marginale. Tutti furono ricoperti da una coltre di foreste e i laghi e i corsi d'acqua furono caratterizzati da una ricca vita acquatica.

Gli agricoltori che giunsero negli altipiani provenendo da sud mossero dapprima verso nordovest seguendo il corso dei fiumi principali dell'Europa centrale: il Danubio, l'Elba, il Meno e il Reno. Essi si insediarono nelle più ricche fra le regioni ecologiche degli altipiani, i bacini fluviali riempiti di löss, dove piante e animali selvatici erano ciononostante relativamente scarsi. Le nuove comunità neolitiche stabilitesi in queste regioni furono perciò costrette, per la loro sussistenza, a fare affidamento su animali e piante da loro introdotti (bovini, pecore, capre e maiali; e orzo e frumento). Col tempo, però, una parte degli immigranti si spinse più a nord lungo la Vistola e l'Oder, penetrando nella pianura baltica.

Quivi essi trovarono un terreno e tipi di suolo considerevolmente diversi da quelli degli altipiani. Il nuovo ecosistema comprendeva più di 450 specie di piante commestibili, pesci, tartarughe e molluschi lamellibranchi nei laghi e nei fiumi, e cervi rossi *(Cervus elaphus)*, caprioli *(Capreolus capreolus)*, bovini e suini selvatici e persino cavalli selvaggi nella foresta. Questa zona di enorme produttività naturale aveva un solo inconveniente: i suoi prodotti non erano costanti nel corso dell'anno e scendevano a livelli esigui in inverno. In che modo i coloni del Neolitico inferiore si adattarono a un ecosistema dalla produttività tanto irregolare è uno dei problemi a cui i siti polacchi stanno fornendo una risposta.

Su un totale di oltre 20 siti neolitici oggi noti in prossimità di Brześć Kujawski, cittadina del voivodato di Wloclawek in Polonia, due (designati come sito n. 3 e sito n. 4) furono scoperti nel 1933 quando alcuni contadini, che stavano asportando ghiaia in prossimità della linea di costa di un antico lago prosciugato, si imbatterono in manufatti e in ossa umane. Konrad Jażdżewski, del Museo archeologico statale di Varsavia, venuto a conoscenza della scoperta, vi trascorse varie stagioni (1933-1939) scavando un'area totale di più di 10 000 metri quadrati. Grazie ai suoi sforzi furono riportati in luce le piante perimetrali di molte case neolitiche, più di 50 sepolture umane, alcune delle quali con ricche suppellettili funerarie formate da ornamenti di rame, osso e perline, e picconi di corna di cervo. Tanto le ossa umane quanto quelle animali si erano ben conservate nel suolo non acido. I manufatti di rame sono fra i più antichi rinvenuti in questa parte dell'Europa centro-settentrionale; il metallo proveniva probabilmente da miniere a sud dei

Il perimetro di una «casa lunga» (o «rettangolare»), venuto in luce dopo la rimozione della parte di suolo, smossa dall'aratura, servì da guida agli scavatori di un sito del Neolitico inferiore a Brześć Kujawski, mentre suddividevano in sezioni il suolo di argilla scolorita in cui erano sorte un tempo strutture verticali in legno. Scoloriture analoghe, nel sottosuolo, segnalarono la presenza di fosse per rifiuti, contenenti manufatti scartati e ossa animali. Lo scavo irregolare in primo piano era l'accesso a uno strato di argilla che i coloni neolitici usavano per «intonacare» i muri delle loro case e per fabbricare ceramiche. Questo è il più recente dei due insediamenti principali di una popolazione della cultura di Lengyel. Datazioni con il carbonio 14 indicano che esso avvenne fra il 4500 e il 4400 a.C. e che la località fu abbandonata nel 3900 a.C.

Ai resti della «casa lunga» illustrata nella foto in alto erano associate tre sepolture: gli scheletri sono quelli di un uomo, a sinistra, coricato sul fianco destro, e di due donne, coricate sul fianco sinistro. I corpi sono orientati con la testa in direzione sud-sudest. Sono state riportate in luce, finora, più di 70 sepolture della cultura di Lengyel, ma non sono state ancora trovate le tombe dei primi abitanti dei siti n. 3 e n. 4 (cultura della *Bandkeramik*). Ricche suppellettili funerarie, tra cui ornamenti in rame, sono a volte associate alle suddette sepolture.

Carpazi. La guerra interruppe queste ricerche e i manufatti scoperti nella stagione di scavi del 1939 andarono distrutti nel bombardamento del deposito ferroviario di Varsavia; Jażdżewski aveva però già delineato l'importanza dei siti per una comprensione della vita del Neolitico inferiore nella pianura baltica.

Tanto il lavoro di Jażdżewski, compiuto prima che venisse introdotto il metodo di datazione con il carbonio 14, quanto ricerche posteriori nella stessa area non furono in grado di chiarire un certo numero di quesiti di natura cronologica ed economica. Di conseguenza, nell'estate del 1976, uno di noi (Grygiel), del Museo di archeologia ed etnografia di Łodź, accompagnato da Bogucki, che si stava allora preparando alla laurea alla Harvard University, si recò a ispezionare i siti n. 3 e n. 4. Trovammo così che aree consistenti di entrambi rimanevano ancora da scavare e, grazie a fondi per ulteriori ricerche, messi a disposizione dal voivodato di Włoclawek e dalla città di Brześć Kujawski, riuscimmo a concludere gli scavi, principalmente nel sito n. 4 in sette successive stagioni.

Come nella maggior parte dei siti neolitici dell'Europa temperata, anche qui i resti archeologici si trovano relativamente in superficie. Lo strato superiore del suolo è costituito da una «terra nera» cuiaviana di peso medio, che poggia su un substrato di argilla e ghiaia. I 30 centimetri superiori vengono smossi regolarmente dall'aratura ogni anno, cosicché il metodo di scavo consiste nell'asportare la zona lavorata con l'aratro, esponendo in tal modo tutti i lineamenti dell'argilla e della ghiaia sterili sottostanti. A differenza dell'accumulo di depositi di varia epoca tipico di un sito stratificato in profondità, qui si trova un palinsesto di cunette, fossati, cavità per pali e tombe, spesso sovrapposti. Tali sovrapposizioni consentono di determinare le età relative di insediamenti successivi. Le età assolute possono essere stabilite mediante datazioni con il carbonio 14; a tutt'oggi in laboratori della Polonia, dei Paesi Bassi e degli Stati Uniti ne sono state determinate in totale quindici.

Nei siti n. 3 e n. 4 sono evidenti due periodi principali di occupazione del Neolitico, separati da un intervallo di due o tre secoli nel corso del quale l'insediamento si ridusse a una scala molto ristretta. Il periodo iniziale si avviò attorno al 5300 a.C. (la vera data secondo una cronologia calibrata sul radiocarbonio) e continuò sino al 4800 a.C. circa. Gli occupanti appartenevano a un gruppo assegnato da archeologi europei alla cultura della *Bandkeramik*, nota anche come «ceramica a nastro» o «ceramica lineare» per le tipiche decorazioni a fasci di linee incise.

I resti di questa cultura si trovano in tutt'Europa dall'Ucraina alla Francia, tanto sugli altipiani ricoperti di löss quanto su parti della pianura dell'Europa centro-settentrionale. Molti fra i siti sugli altipiani presentano tracce di solide «case lunghe» (o «rettangolari»), a volte solo una singola «fattoria», ma più spesso varie abitazioni raggruppate in un villaggio. Nessuna «casa lunga» della cultura della *Bandkeramik* è stata, invece, trovata nella pianura baltica, e i due siti da noi scavati non fanno eccezione. Fra le fosse per rifiuti, grandi ma poco profonde, che segnalavano la presenza della *Bandkeramik*, c'erano solo poche cavità per pali disseminate qua e là, le quali indicavano probabilmente la localizzazione in quei punti di capannoni o di ripari temporanei analoghi. Si sono potute distinguere due fasi separate di occupazione della cultura della *Bandkeramik*. La fase anteriore si trovava sulla punta di una penisola sulla riva di un lago nel sito n. 3 e quella posteriore era a circa 200 metri di distanza nel sito n. 4. In nessuno dei due siti furono trovate sepolture tipiche della cultura della *Bandkeramik*.

La popolazione della cultura di Lengyel costruì con corna di cervo vari oggetti, tra cui picconi e cucchiai come quello della foto in alto, trovato in una delle molte fosse per rifiuti del sito n. 4 e primo manufatto del genere rinvenuto fra i resti del Neolitico nell'Europa centro-settentrionale. Della stessa cultura sono state anche trovate, fra le suppellettili funerarie, perline fatte di valve di lamellibranchi infilate come quelle della foto qui sopra, associate a uno scheletro di donna.

Vari secoli dopo, emigranti assegnati da archeologi europei alla cultura di Lengyel occuparono entrambi i siti. Questa manifestazione del Neolitico inferiore, che prende il nome dal sito d'origine in Ungheria, rappresenta una prosecuzione essenziale, nell'Europa centro-settentrionale, delle tradizioni della *Bandkeramik*, presenti nell'Europa orientale, ma è caratterizzata da nuove forme di ceramica e da un declino nella quantità di decorazioni sul vasellame. Anche molti siti della cultura di Lengyel hanno «case lunghe» a pianta trapezoidale, più strette a un'estremità e case lunghe di questa forma tipica sono state riportate in luce sia sugli altipiani di löss sia nella pianura baltica: esse costituiscono i primi ripari permanenti che compaiono nei due siti di Brześć Kujawski.

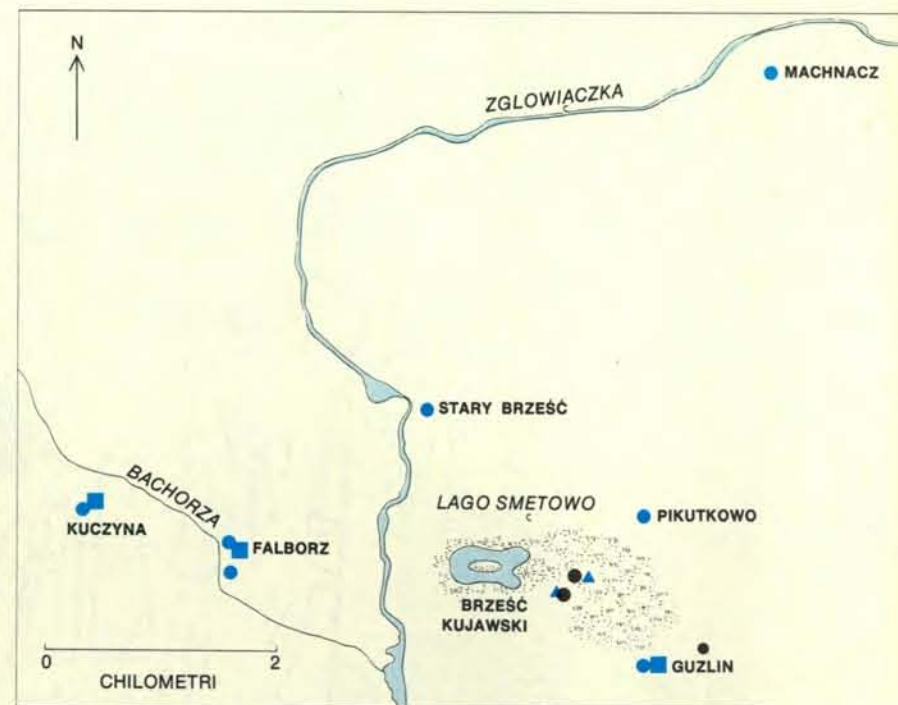
Nei siti possono essere identificate quattro fasi di Lengyel. La più antica, evidente soprattutto nel sito n. 4, apparve fra il 4500 e il 4400 a.C. e continuò sino al 4300 a.C. circa. Le tre posteriori, che si estesero dal 4300 a.C. circa sino al 3900 a.C. circa, sono evidenti in entrambi i siti. Negli anni trenta Jażdżewski chiamò queste forme locali posteriori «gruppo di Brześć Kujawski», denominazione ancora in uso.

I segni più sorprendenti fra quelli trovati scavando nel sottosuolo dei due siti sono le tracce di case lunghe della cultura di Lengyel. Essi appaiono come lunghe superfici scure scolorite nell'argilla e sono i resti di cavità dove venivano piantati i pali verticali delle pareti delle case. Fra l'inizio dei lavori, negli anni trenta, e oggi, sono state scoperte più di 50 «piante di case», alcune delle quali lunghe più di 30 metri e orientate tutte con l'asse maggiore che corre da nordovest (l'estremo più stretto) verso sudest (l'estremo più largo). In media le case della cultura di Lengyel erano lunghe una ventina di metri; il lato orientato verso nordovest era largo circa tre metri, e quello verso sudest circa cinque metri.

La ragione per cui la popolazione della cultura di Lengyel costruiva abitazioni trapezoidali rimane incerta, ma una spiegazione probabile è che l'estremità più stretta fosse rivolta verso i venti prevalenti. Nel 1976 una copia di una «casa lunga» simmetrica del Neolitico fu costruita in Francia come sostegno di un'antenna televisiva. L'estremità orientata verso i venti prevalenti ha già sofferto gravi danni in conseguenza delle intemperie, mentre gli altri tre muri sono ancora in buone condizioni. Nei siti n. 3 e n. 4 vi sono chiari indizi che queste «case lunghe» sono state ricostruite varie volte; la ripetuta riutilizzazione degli stessi siti fa pensare che l'occupazione sia stata continua in tutte le quattro fasi, probabilmente con anche una decina di case occupate simultaneamente.

I segni più comuni e più ricchi di informazioni ritrovati al di sotto della superficie, nei due siti, sono fosse di varia forma e grandezza. Come abbiamo già osservato, quelle in cui gli uomini della cultura della *Bandkeramik* gettavano i rifiuti sono generalmente grandi e poco profonde. Al tempo degli insediamenti, può darsi che non fossero altro che depressioni naturali nel terreno, paragonabili all'avvallamento che rimane nel suolo quando viene sradicato un albero caduto. Esse contengono densi cumuli di detriti, formati da ossa di animali e da manufatti in pietra, in selce e in ceramica. Molte fosse della cultura di Lengyel sono, invece, profonde e irregolari, scavate in parti del sito in cui immediatamente sotto la superficie c'è argilla. Ciò fa pensare che siano state scavate in origine per procurarsi argilla con cui «intonacare» le proprie abitazioni di legno e per produrre vasellame.

Altre fosse di Lengyel sono in generale profonde e circolari e si trovano presso la riva dell'antico lago oggi prosciugato. Depositi di fango sul fondo danno motivo di pensare che esse contenessero talvolta acqua stagnante, probabilmente per conservarvi lamellibranchi, tartarughe e pesci vivi. In ogni caso, alla fine, la maggior parte delle fosse da cui era stata ricavata argilla e delle fosse-vivai è stata utilizzata

I siti di Brześć Kujawski sono 150 chilometri a ovest-nordovest di Varsavia e circa 200 chilometri a nordovest dei Monti della Santa Croce, un'area in cui si trova la selce cioccolato, così chiamata per il suo colore. Questa selce era molto apprezzata dalla popolazione della cultura della *Bandkeramik* per le proprietà rivelate nella scheggiatura. I gruppi umani che occuparono in seguito quei siti si accontentarono della selce di qualità inferiore depositata dai ghiacciai.

Le occupazioni successive a Brześć Kujawski si localizzarono su una penisola sul Lago Smetowo, allora molto più grande di oggi. Gli insediamenti della popolazione della *Bandkeramik* nei siti n. 3 e n. 4 erano in prossimità della punta della penisola e della sua base *(dischetti in nero)*; gli insediamenti della cultura di Lengyel, negli stessi due siti, sono segnati da triangoli in colore. Altri siti della cultura della *Bandkeramik (dischetti in colore)* sono stati trovati a sud, a ovest e a nord. Sono stati trovati anche tre siti non residenziali della cultura di Lengyel *(quadratini in colore)*. Uno di essi, a giudicare dalle molte asce portate in luce doveva servire per il taglio di legname.

per gettarvi rifiuti. A differenza delle fosse della cultura della *Bandkeramik*, quelle della cultura di Lengyel presentano variazioni considerevoli nel contenuto dei rifiuti. Le ossa di animali e i manufatti gettati via sono più numerosi in quelle vicino alle case lunghe.

In alcune aree in cui il sottosuolo fu smosso in epoca neolitica, lo scolorimento è minore di quanto non sia nelle fosse riempite con rifiuti; a quanto pare queste aree scavate erano state poi ricoperte di terriccio senza alcuna mescolanza con detriti organici. Studiate dagli archeologi, esse si rivelarono di solito tombe della cultura di Lengyel. A tutt'oggi è stata riportata in luce oltre una settantina di queste sepolture, la maggior parte delle quali, negli anni trenta, per opera di Jażdżewski. In questa sorta di tombe individuali sono stati trovati, in genere, corpi umani. Di norma, gli scheletri sono in posizione contratta con le ginocchia piegate; gli individui di sesso maschile venivano sepolti sul fianco destro, e quelli di sesso femminile sul fianco sinistro. L'orientamento è lo stesso per entrambi i sessi: i crani poggiano in prossimità del margine meridionale della fossa rivolti a sudest. Le donne portavano spesso perle e altri ornamenti di conchiglie, osso e rame; agli uomini erano spesso associati picconi di corna di cervo. Finora non sono state trovate sepolture della cultura della *Bandkeramik*.

Nel loro insieme i manufatti lasciati dalle popolazioni di entrambe le culture comprendono tipicamente asce e accette neolitiche fatte di pietra «scheggiata» e levigata. Un'ascia di questo tipo (simile a una forma per calzature e detta *celt*) è tipica della cultura della *Bandkeramik*. Questi antichi agricoltori producevano al tempo stesso altri piccoli utensili scheggiando selce. La lavorazione della selce nelle due culture era però del tutto diversa. La popolazione della *Bandkeramik* privilegiava una varietà polacca: la selce cioccolato, così chiamata per il suo colore marrone, la quale si trova solo in una regione della Piccola Polonia, i Monti della Santa Croce, 200 chilometri circa a sudest dei siti esaminati. La selce cioccolato polacca era già apprezzata dai produttori di utensili litici in epoca paleolitica: essa si prestava in modo eccellente a essere scheggiata e la popolazione della *Bandkeramik* ricavava dai nuclei di questa selce lame lunghe e sottili.

La popolazione della cultura di Lengyel non manifestò una preferenza così spiccata. Essa produsse infatti la maggior parte dei suoi utensili di schegge con gli «erratici» di selce che gli strati di ghiaccio, ritirandosi, avevano lasciato generosamente disseminati qua e là nella pianura nordeuropea. Gli erratici di selce molti dei quali sottoposti a pressioni e ad altri tipi di sollecitazioni nel corso del loro trasporto per opera dei ghiacciai tabulari, tendono a essere scarsamente soggetti alla scheggiatura, e le schegge di selce della cultura di Lengyel mostrano questa prerogativa. Solo quando gli uomini di Lengyel avevano bisogno di un utensile ben affilato, di grande precisione, si davano la pena di procurarsi e di lavorare una selce cioccolato.

La popolazione di Lengyel andò un passo oltre rispetto a quella della *Bandkeramik*: produsse, infatti, asce di pietra scheggiata e levigata con un foro, in modo da potervi inserire un manico, e i pezzi di pietra tolti dal nucleo nel corso di questa lavorazione sono stati ritrovati nelle fosse dei rifiuti. La popolazione di Lengyel utilizzava, inoltre, le corna caduche dei cervi rossi (animali che le sostituiscono ogni anno) sia per produrre picconi sia come materia prima per altri manufatti. Particolarmente notevole fra questi un bel cucchiaio di corno, unico nell'Europa centro-settentrionale, che abbiamo scoperto solo l'anno scorso.

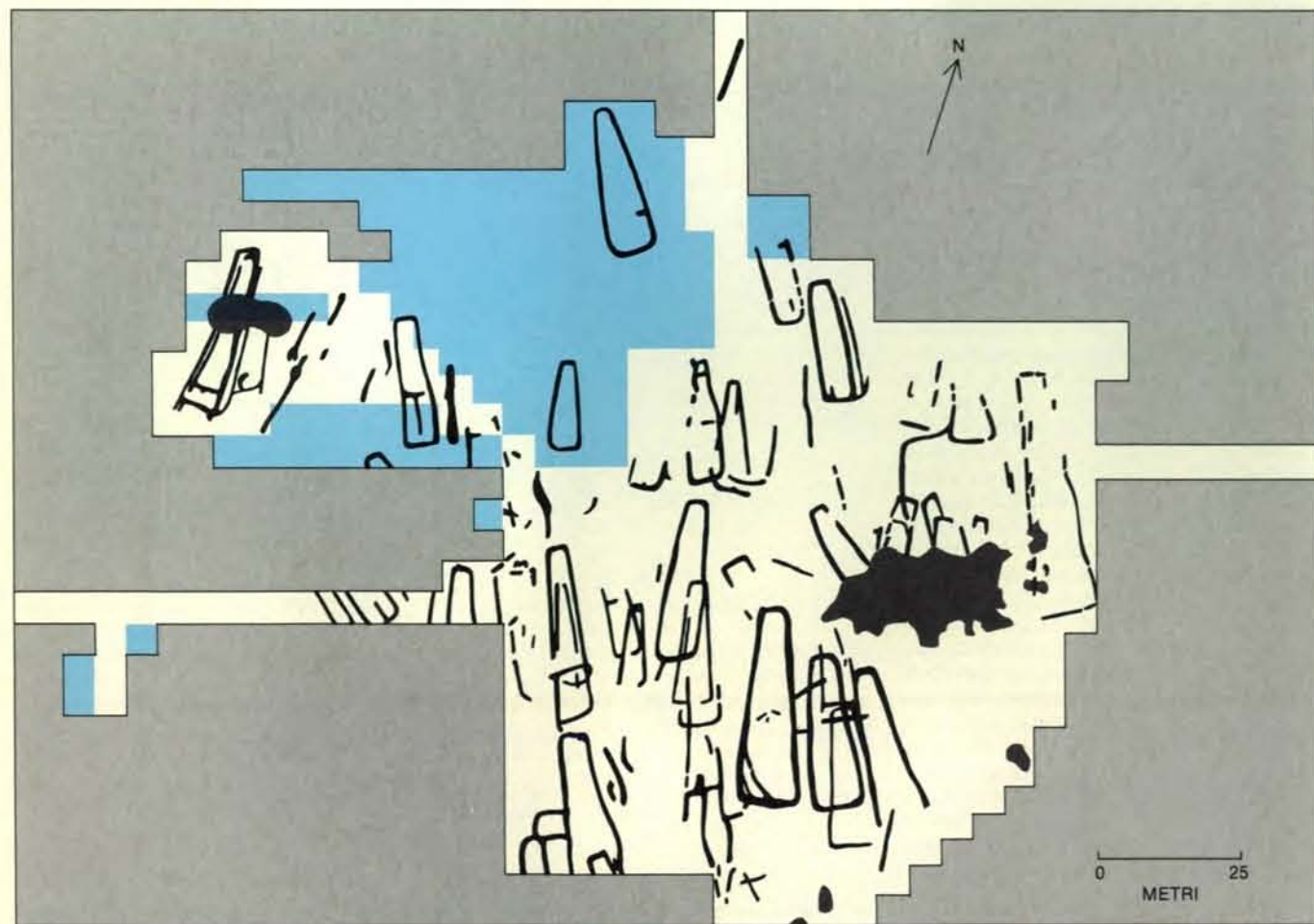
Ossa e denti di animali venivano utilizzati anche per produrre perle per collane e pendenti. Le coste di mammiferi di grandi dimensioni, probabilmente bovini, venivano incurvate a formare braccialetti, a quanto pare mentre erano ancora «fresche», e poi venivano arricchite con incisioni. Grandi quantità di perline si ottenevano dalle valve madreperlacee di *Unio*, un lamellibranco d'acqua dolce. Attorno alla vita di una donna, in una tomba, c'erano vari fili formati da più di 5000 perline di *Unio*, ciascuna delle quali del diametro di circa un centimetro. Le tombe di donne della cultura di Lengyel contengono anche perline di rame e amuleti; strisce e sbarrette di rame trovate nelle fosse dei rifiuti indicano che questi ornamenti venivano fatti sul posto con pezzi di rame importato.

La ceramica da cui la cultura della *Bandkeramik* prende il nome differisce da quella di Lengyel non solo per motivi decorativi, ma anche per forma. A Brześć Kujawski la forma più comune è quella stessa che si incontra molto spesso in altri siti della *Bandkeramik* nell'Europa centrale: un vaso globulare che forma tre quarti di una sfera. Un'altra forma della *Bandkeramik*, comune nei due siti polacchi e altrove nel bassopiano dell'Europa centro-settentrionale, ma relativamente rara sugli altipiani, è un setaccio. Non abbiamo trovato setacci intatti, ma il gran numero di frammenti rinvenuti fa pensare che questi manufatti servissero a separare la crema dal siero di latte nella produzione di formaggio. Robert Cowie, dell'Università di Sheffield, sta attualmente compiendo studi su cocci di vasi della *Bandkeramik* per vedere se la loro composizione minerale offra indizi circa l'origine dell'argilla usata nella loro produzione.

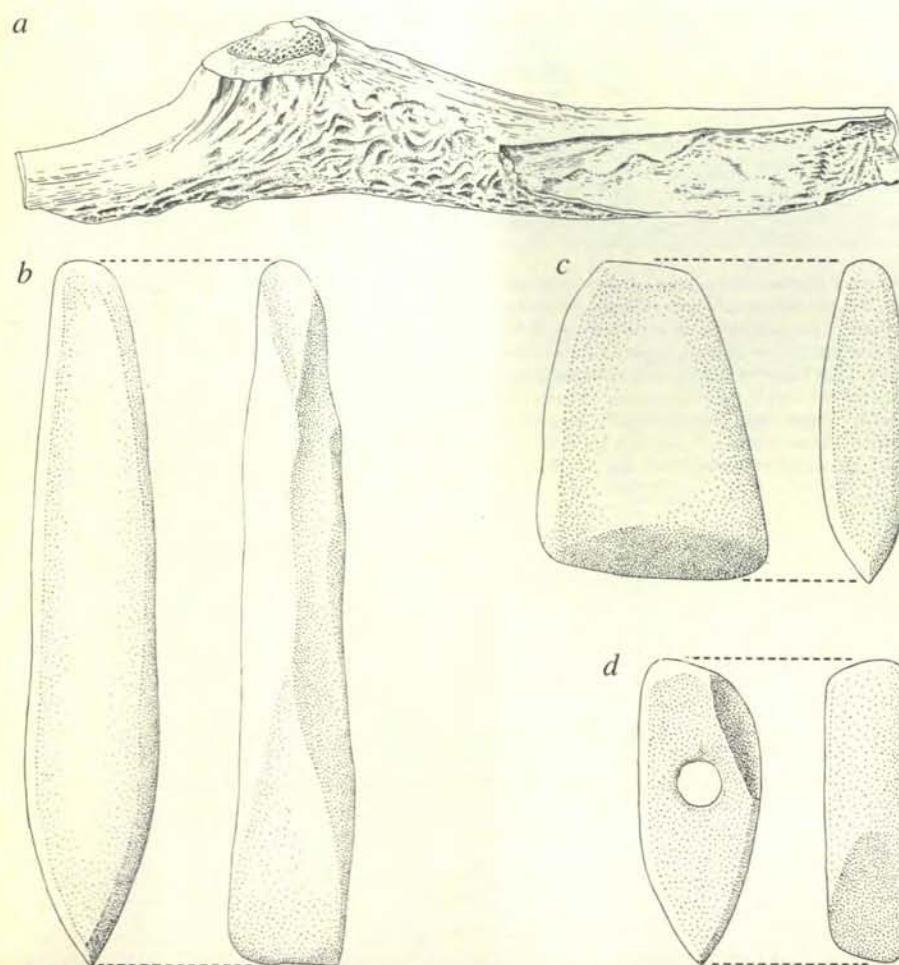
Le ceramiche della cultura di Lengyel vanno da ciotole poco profonde a vasi simili ad anfore greche, e comprendono oggetti specializzati come cucchiai di ceramica. Non è stato ancora determinato con precisione il contenuto in minerali, ma la grande quantità di mica presente nelle ceramiche fa pensare che siano state prodotte con materiali disponibili in loco. La stessa conclusione viene suggerita dal fatto che dalle fosse presenti nel sito è stata estratta dell'argilla.

Nel corso delle nostre ricerche, a partire dal 1976, abbiamo recuperato una grande quantità di ossa animali in buone o eccellenti condizioni di conservazione. Per esempio, abbiamo separato più di 6000 lische di pesce, per lo più di pesce persico e abramide. Anche a questo proposito si nota un contrasto fra i resti di animali tipici degli insediamenti anteriori e quelli degli insediamenti posteriori. I principali animali domestici, come si è riscontrato anche per altri siti della cultura della *Bandkeramik* nel bassopiano dell'Europa centro-settentrionale, erano bovini. Poiché, nell'ambito delle dimensioni delle ossa dei bovini, la maggior parte delle ossa ritrovate si colloca a livello inferiore, pare che le mandrie fossero composte principalmente da mucche. La popolazione della *Bandkeramik* allevava anche pecore e capre; negli insediamenti sugli altipiani, teneva anche molti maiali, ma purtroppo per l'intero periodo di 500 anni di occupazione da parte di questa popolazione abbiamo potuto riportare in luce a Brześć Kujawski tracce di soli tre maiali. Altrettanto rimarchevole è la scarsità di resti di uccelli, pesci, tartarughe e molluschi lamellibranchi e di ossa di cervo rosso, capriolo e maiale selvatico. Del gran numero di lische di pesce trovate nei due siti, meno di 100 provengono dalla popolazione della *Bandkeramik* che li ha occupati.

Le ossa di animali associate all'occupazione da parte della popolazione della cultura di Lengyel ci presentano un quadro diverso. Benché nel complesso i resti di bovini di questa popolazione siano in qualche misura più numerosi di quelli della popolazione della *Bandkeramik*, essi non figurano con lo stesso rilievo nel totale dei resti di animali. Pecore e capre sono infatti altrettanto numerose, e i maiali uguagliano in numero i bovini. Contemporaneamente si registra un aumento nel numero di resti di animali selvatici: cervi rossi e caprioli, uccelli acquatici, tartarughe, pesci e molluschi lamellibranchi. Non tutti gli animali cacciati erano usati come cibo: i castori, per esempio, venivano evidentemente uccisi per utilizzarne la pelle, e lo stesso vale forse per le lontre, le donnole e le lepri.

Che cosa si può dire sulla pratica dell'agricoltura in questi primi insediamenti del Neolitico inferiore? Noi abbiamo dedicato particolare attenzione al recupero di materiali vegetali per mezzo di tecniche di setacciamento e galleggiamento di materiali immersi in acqua. Nelle fosse di rifiu-

Konrad Jażdżewski, del Museo archeologico statale di Varsavia, riportò in luce fra il 1933 e il 1939 nel sito n. 4 le tracce delle pareti di più di 50 «case lunghe», molte delle quali sovrapposte. I profili nell'area bianca illustrano il lavoro compiuto negli anni trenta. Le aree in grigio scuro indicano interventi in epoca moderna, prima dell'inizio dei lavori di Jażdżewski. Le aree in colore sono state scavate dagli autori fra il 1976 e il 1982. Sono state omesse dalla pianta sepolture e fosse per rifiuti *(si veda l'illustrazione nella pagina a fronte)*.

I lavori di scavo compiuti da Jażdżewski nel sito n. 3, nella parte più vicina alla riva del lago, sono inclusi all'interno dell'area colorata. In essa sono compresi i profili, quasi perfettamente sovrapposti, di due «case lunghe» e un certo numero di fosse per rifiuti similmente sovrapposte, che contenevano manufatti e resti animali appartenenti a due occupazioni, della cultura della *Bandkeramik* e della cultura di Lengyel. È evidente la forma trapezoidale delle «case lunghe» di Lengyel; l'estremità più stretta era orientata forse verso il vento prevalente. Altrove in Europa si trovano «case lunghe» della cultura della *Bandkeramik*, mentre non se ne conosce alcuna nel bassopiano dell'Europa centro-settentrionale. Le cavità per pali associate ai manufatti della *Bandkeramik* fanno pensare che questi antichi migratori del Neolitico abbiano costruito qui solo ripari di tipo temporaneo.

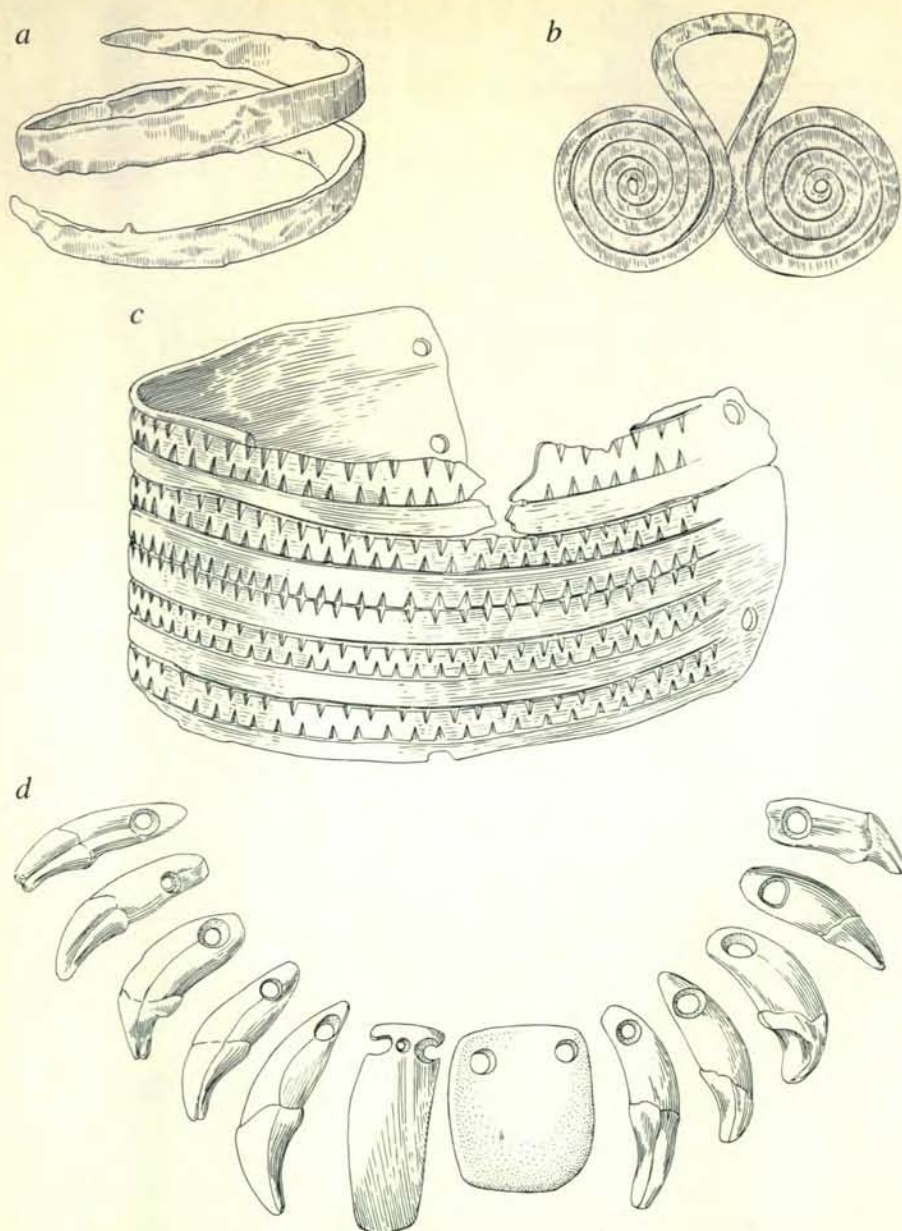
Il vaso globulare di ceramica, in forma di tre quarti di sfera *(a)*, illustra la decorazione, formata da fasci di linee, che dà alla cultura della *Bandkeramik* o «Ceramica a nastro» il suo nome. Sotto il vaso c'è parte di un setaccio *(b)*, un manufatto comune della *Bandkeramik* nei siti polacchi, anche se non sono stati trovati setacci interi. Essi servivano probabilmente a separare la crema dal siero di latte nella produzione di formaggi. La pentola con manici dalla decorazione semplice *(c)* è una forma tipica della cultura di Lengyel. Al di sotto *(d)* c'è un altro oggetto comune della manifattura di Lengyel: un cucchiaio in ceramica. Setaccio e cucchiaio non sono in scala.

Le asce delle popolazioni delle culture della *Bandkeramik* e di Lengyel erano fatte con pietra scheggiata e levigata o con corna di cervo rosso *(a)*. L'ascia in pietra a sinistra *(b)*, nota come *celt*, e simile a una forma per calzature, è un utensile tipico della cultura della *Bandkeramik*, mentre l'ascia più corta e più larga in *(c)* è tipica della cultura di Lengyel. La popolazione di Lengyel praticava anche fori in alcune asce *(d)* per inserirvi un manico. Nessuna ascia con foro è stata trovata intatta nei siti n. 3 e n. 4, mentre nelle fosse per rifiuti sono stati trovati i pezzi di pietra asportati nel praticare i fori.

ti della cultura della *Bandkeramik* siamo riusciti a trovare solo tre chicchi di cereali carbonizzati; ciò fa pensare che questa popolazione si dedicasse molto poco alla coltivazione di cereali. Le fosse dei rifiuti della cultura di Lengyel, di contro, hanno fornito grandi quantità di cereali carbonizzati. Una fossa di cui abbiamo recuperato i contenuti lo scorso anno ha fornito più di 300 esemplari. Un'analisi preliminare eseguita da Caroline Quillian Stubbs, che lavorava allora all'Università di Tübingen, indica che la maggior parte dei cereali di Lengyel è costituita dalla forma antica di frumento noto ai botanici come farro *(Triticum dicoccum)*.

Non è chiaro se la popolazione della cultura di Lengyel abbattesse foreste per ampliare le proprie attività agricole. Per poter giungere a tale conclusione dovremmo prima analizzare i granuli di polline a varie profondità nel suolo. A tutt'oggi un tale profilo dei pollini non è ancora disponibile per alcuno dei due siti, anche se le ricerche in proposito dovrebbero arrivare presto alla loro conclusione. Una possibile indicazione di attività di diboscamento è fornita però dalla presenza, fra i resti di animali, di ossa di aquile e di falchi di palude. Entrambe queste specie di rapaci necessitano di habitat caratterizzati da spazi aperti.

Per determinare quali risorse le popolazioni del Neolitico inferiore sfruttassero nelle varie stagioni dell'anno, ci siamo basati su numerosi indizi diversi, fra cui denti di maiali e di cervi e il grado di eruzione dei denti ancora presenti nelle mascelle degli animali. Abbiamo preso nota anche di altre indicazioni di stagionalità, come la presenza, fra le ossa di animali, dei resti di certe specie di uccelli acquatici (indicativi di attività di caccia invernale e primaverile), di resti di tartarughe d'acqua dolce (animali inattivi dal tardo autunno all'inizio della primavera) e di corna caduche di cervi (i cervi perdono le corna ogni anno, in febbraio e in marzo, per poi riformarle). Nei depositi di rifiuti della cultura della *Bandkeramik* non siamo riusciti a trovare alcuna indicazione di attività invernali. Al tempo stesso, la presenza in essi di ossa di tartaruga indica che il gruppo della *Bandkeramik* era attivo in questi siti nei mesi estivi. I contenuti dei depositi di rifiuti della cultura di Lengyel dimostrano, invece, che questa popolazione che occupò posteriormente i siti vi viveva per tutto l'anno.

Particolarmente ricco di informazioni si rivelò l'esame degli «anelli di crescita» nei denti di mammiferi. Questi strati di cemento si depositano sotto forma di bande opache d'estate e di bande traslucide d'inverno. Sezionando un dente e determinando a quale tipo appartenga lo strato depositato per ultimo, si può accertare grosso modo in quale periodo dell'anno l'animale sia morto. In questo modo abbiamo stabilito che la popolazione della cultura di Lengyel di solito macellava i maiali alla metà dell'inverno e cacciava cervi rossi e caprioli verso la fine dell'inverno e in primavera.

Oltre alle ricerche compiute nei siti n. 3 e

Fra gli ornamenti tipici delle tombe della cultura di Lengyel vi sono un bracciale in rame *(a)* e un pendente pure in rame in forma di *pince-nez (b)*. Bracciali d'osso incisi *(c)* venivano ottenuti da coste probabilmente di bovini. In *(d)* collane con pendenti di corallo fossile levigato e denti di cane o di lupo, forati per passare il filo. I manufatti in rame, sorprendenti in un contesto del Neolitico inferiore, furono eseguiti probabilmente sul posto con rame trasportato dal sud dei Carpazi.

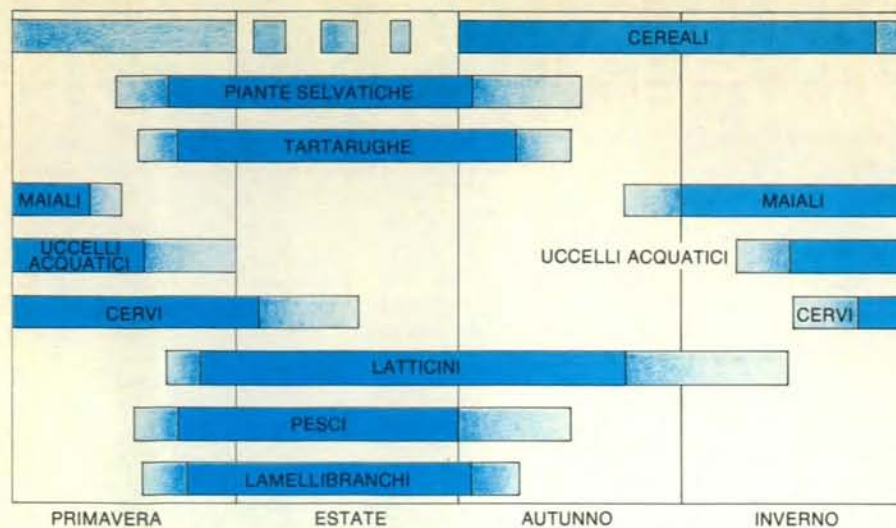
n. 4, abbiamo perlustrato le campagne circostanti alla ricerca di altri siti neolitici e abbiamo trovato tracce di altre sei occupazioni attribuibili alla cultura della *Bandkeramik*. Esse presentavano tutte le stesse caratteristiche dei siti principali nei pressi di Brześć Kujawski: mancavano indizi che potessero far pensare ad abitazioni permanenti e le fosse in cui venivano accumulati i rifiuti erano poco profonde. Sono stati scoperti anche altri tre siti della cultura di Lengyel, tutti entro un'ora di cammino dai due siti principali. Anche qui mancava ogni indizio di abitazioni permanenti e le poche fosse di rifiuti hanno fornito manufatti di varietà più limitata rispetto a quelle nei siti principali. Si ha l'impressione che in queste aree dovessero svolgersi attività specializzate, forse pratiche agrarie o di cura del bestiame. Pare che uno di questi siti, a Kuczyna, fosse un campo per il taglio del legname: nelle sue fosse per rifiuti sono state trovate in gran numero teste d'ascia in pietra e in corno di cervo.

In tutti questi casi, sia la vicinanza degli accampamenti della popolazione della cultura di Lengyel all'area principale di insediamento sia la loro scarsità di strutture permanenti fanno pensare che gli agricoltori di Lengyel occupassero un «territorio» ben definito, mentre la maggiore dispersione degli accampamenti della popolazione della *Bandkeramik* dà l'impressione opposta. Questo, insieme alla differenziazione funzionale dei campi satelliti di Lengyel, indica che gli agricoltori di questa cultura sfruttavano l'ambiente circostante in modo più organizzato rispetto ai loro precedessori della cultura della *Bandkeramik*.

Quali conclusioni generali si possono trarre dalle ricerche svolte a Brześć Kujawski? In passato tanto l'economia di sussistenza quanto la cultura materiale delle prime comunità neolitiche dell'Europa centro-settentrionale sono state spesso implicitamente considerate «in blocco». Per esempio, ogni volta che si trovano oggetti appartenenti alla *Bandkeramik*, si presume che si troverà l'intero complesso di elementi formato da «case lunghe», asce simili a forme per calzature, cereali, ossa di bovini, pecore, capre e maiali. Nei due siti polacchi non è però così, e neppure in altri siti della *Bandkeramik* localizzati nel bassopiano dell'Europa centro-settentrionale. Le «case lunghe» degli altipiani sono assenti, le tracce di coltivazione di cereali sono sparse e le ossa di maiale sono una rarità. Inoltre le occupazioni dei siti n. 3 e n. 4 da parte di popolazioni della *Bandkeramik* non presentano alcuna indicazione di un soggiorno invernale. Vale invece l'opposto per le occupazioni degli stessi siti da parte di uomini della cultura di Lengyel e anche per i siti della cultura della *Bandkeramik* nella zona di altopiano più a sud.

Queste informazioni esigono di riconsiderare il modo di vita delle più antiche comunità produttrici di cibo della pianura baltica. Le omissioni in cui si è incorsi nel considerare il «blocco» sopracitato e l'alta proporzione di ossa di bovini fra i resti animali della cultura della *Bandkeramik* ci fanno pensare che questa prima escursione neolitica a nord degli altipiani non possa essere interpretata come formazione di insediamenti permanenti di agricoltori-allevatori, ma vada vista piuttosto come transumanze stagionali di bovini (con produzione di formaggi) nelle foreste dei bassipiani. La vegetazione ecologicamente diversa dei bassipiani deve aver fornito un ricco foraggio estivo per bovini nonostante la grande quantità di alberi presenti, un foraggio probabilmente molto più ricco di quello dei pascoli disponibili negli angusti fondivalle degli altipiani.

Le condizioni propizie alle migrazioni stagionali del bestiame potrebbero essere state ancor più favorite in quest'area forestale da una precedente attività degli abitanti indigeni del Mesolitico (preneolitici): il deliberato diboscamento di aree forestali con il fuoco. La prova che questo diboscamento risalga al Mesolitico è fornita dalla proporzione anormalmente elevata di polline di nocciolo nei profili pollinici che rappresentano il periodo datato al 5800 a.C. circa, mezzo millennio prima che le popolazioni della cultura della *Bandkeramik* entrassero per la prima volta nei bassipiani. Il nocciolo è una pianta che cresce di solito su terreni in precedenza diboscati.

Una prova ulteriore dell'importanza dei latticini per questi primi immigranti neolitici è la presenza di setacci d'argilla sia a Brześć Kujawski sia in altri siti della cultura della *Bandkeramik* nelle pianure dell'Europa centro-settentrionale; questi

Il ciclo alimentare annuale della popolazione della cultura di Lengyel, a cominciare dalla fine dell'inverno (*a sinistra*), faceva risaltare la dipendenza da un unico animale domestico (il maiale) e da tre animali selvatici (un uccello acquatico migratore, il cervo rosso e il capriolo) in attesa che si rendessero disponibili, verso la fine della primavera e fino all'autunno, cibi vegetali forniti da piante selvatiche, molluschi lamellibranchi, tartarughe e pesci d'acqua dolce, assieme al latte prodotto da mucche, pecore e capre. Col raccolto del frumento in autunno, la popolazione della cultura di Lengyel acquisiva una risorsa importante sino all'inverno inoltrato, quando subentravano come altra fonte principale di cibo i maiali ingrassati durante l'estate, integrati da uccelli acquatici e cervi e dal formaggio, un prodotto del latte che poteva essere conservato a lungo.

manufatti sono relativamente rari nei siti della cultura della *Bandkeramik* sugli altipiani. Si può presumere che una fra le ragioni di questa scarsità risieda nel fatto che i coloni degli altipiani avevano a disposizione non solo bovini, pecore e capre, ma anche maiali e cereali, mentre la dieta degli abitanti dei bassipiani doveva essere limitata a cibi di origine vegetale e a latticini. Un'altra ragione potrebbe consistere nel fatto che le popolazioni neolitiche della cultura della *Bandkeramik* nell'Europa centro-settentrionale non dovevano ancora avere sviluppato la tolleranza al lattosio (zucchero del latte) tipica dei popoli europei moderni; perciò questi uomini avevano bisogno di lavorare il latte di mucca trasformandolo in formaggio (e forse in yogurt) per renderlo digeribile. Sulla base degli indizi disponibili, pare probabile che gli allevatori di bestiame della cultura della *Bandkeramik* migrassero a nord con i loro animali nella stagione calda, ritornando però d'inverno ai loro insediamenti permanenti sugli altipiani.

Gli agricoltori-allevatori della cultura di Lengyel che si insediarono sugli altipiani dopo le popolazioni della cultura della *Bandkeramik* lasciarono tracce di diversificazione delle risorse nel corso dell'anno, del tipo che ci si potrebbe attendere da una economia agricola affermatasi di recente. Benché essi coltivassero cereali, è improbabile che il raccolto di farro fosse sufficiente a nutrire la comunità fino al raccolto successivo e che al tempo stesso potesse fornire sementi per la semina primaverile. Questi uomini dovevano certamente raccogliere prodotti spontanei come nocciole e chenopodi. Il chenopodio viene consumato sia come verdura sia come semi e trovò largo impiego anche altrove in Europa durante il Neolitico; in questa parte della Polonia esso raggiunge anche un'altezza di un metro e mezzo. La popolazione di Lengyel potrebbe anche avere integrato i cereali con legumi, come piselli e lenticchie; entrambi si trovano in altri siti neolitici dell'Europa centro-settentrionale, benché nessun indizio della loro presenza sia emerso finora a Brześć Kujawski.

Importanza ancora maggiore nella dieta della cultura di Lengyel avevano gli animali domestici e la cacciagione. Come ci consentono di affermare i nostri studi sui denti, la popolazione della cultura di Lengyel uccideva i maiali d'inverno, dopo averli ingrassati nel corso dell'estate, e verso la fine dell'inverno e in primavera si procurava altra carne dando la caccia al cervo rosso e al capriolo. Anche tartarughe, pesci e molluschi lamellibranchi devono avere dato un contributo alla loro dieta in primavera e in estate, integrando i cibi vegetali ricavati dalla foresta circostante e il latte di mucca, pecora e capra. Una tale combinazione di prodotti selvatici e domestici era verosimilmente molto gradita verso la fine dell'estate, quando il raccolto e l'immagazzinamento del farro dovevano richiedere una grande quantità di lavoro in poco tempo.

L'organizzazione e programmazione di questo tipo di sistema di sussistenza rese evidentemente necessario un mutamento, al tempo della cultura di Lengyel, del genere di insediamenti sparsi proprio della cultura della *Bandkeramik*. Abbiamo già menzionato i campi satelliti della cultura di Lengyel che contribuirono alla conservazione dell'area residenziale permanente; ma ci fu anche, probabilmente, un certo numero di altri luoghi - stazioni di caccia e posti avanzati raggiunti dalle mandrie transumanti - di cui non si trova traccia nei reperti archeologici. Gli agricoltori della cultura di Lengyel dovevano conoscere molto bene i boschi, i campi e le paludi circostanti, e ogni metro quadrato dava un contributo alla prosperità dell'insediamento centrale.

La lunga continuità dell'occupazione da parte della popolazione della cultura di Lengyel attesta il successo con cui questi agricoltori del Neolitico sfruttarono l'ecosistema della foresta circostante. Ogni ecosistema, però, è elastico solo fino a un certo punto. Pare probabile che, proprio al graduale deterioramento dell'ambiente, dopo centinaia di anni di occupazione umana, si debba l'abbandono dei siti di Brześć Kujawski attorno al 3900 a.C. A quell'epoca, però, la produzione di cibo si era affermata stabilmente come forma di vita nel bassopiano dell'Europa centro-settentrionale. La Rivoluzione neolitica stava allora diffondendosi lungo la costa del Baltico, nella Germania settentrionale e in Danimarca, per opera di popolazioni posteriori, come la popolazione della cultura del «bicchiere con collo a imbuto». Le popolazioni antesignane, appartenenti alle culture della *Bandkeramik* e di Lengyel, erano ormai scomparse, ma l'agricoltura, con i suoi innumerevoli effetti, si era ormai affermata stabilmente nell'Europa temperata.

La fossa del sito n. 3, in prossimità dell'antica riva del lago, appare qui in sezione trasversale semplificata. La presenza di sabbia intrusiva e di un sottile strato di valve di molluschi lamellibranchi fa pensare che fosse usata come riserva per molluschi, tartarughe e pesci vivi.

Fisica intuitiva

Le leggi newtoniane del moto sono ben note, ma molti credono che i corpi in movimento si comportino diversamente e, in pratica, seguano una teoria ritenuta valida nei tre secoli prima di Newton

di Michael McCloskey

Si immagini una persona che cammina di buon passo tenendo una pietra all'altezza della spalla. Che cosa accadrà se la persona lascia cadere la pietra? Che tipo di percorso questa seguirà? Molti, a queste domande, rispondono che la pietra cadrà verticalmente, toccando terra esattamente sotto il punto dove è stata lasciata. Alcuni sono addirittura convinti che la pietra si muoverà all'indietro, per atterrare prima del punto in cui è stata lasciata libera. In realtà, il suo moto sarà contemporaneamente in avanti e verso il basso: toccherà terra qualche decina di centimetri davanti al punto di inizio della caduta. La meccanica newtoniana spiega che, quando viene lasciata libera, la pietra prosegue il suo moto in avanti alla stessa velocità della persona che cammina, perché (se si ignora la resistenza dell'aria) nessuna forza influisce sulla componente orizzontale della velocità. Mentre la pietra prosegue in avanti, cade con una velocità che aumenta costantemente. I moti in avanti e verso il basso si combinano: il risultato è una traiettoria di forma molto simile a una parabola.

La discrepanza fra l'intuizione e la realtà fisica è, in questi casi, sorprendente. Ci si potrebbe aspettare, come risultato dell'esperienza quotidiana, che la gente abbia idee abbastanza precise sul moto di oggetti in situazioni familiari. Si potrebbe anche supporre che la gente abbia una vaga intuizione dei principi generali che governano il moto. Invece, sembra che le cose non stiano così: alcuni studi recenti sulla natura, sullo sviluppo e sull'applicazione delle conoscenze relative al moto, indicano che molte persone compiono errori stupefacenti nell'interpretare il moto di oggetti in circostanze molto semplici. Tali errori sembrano essere legati a una teoria del moto intuitiva, ma sistematica, che è in contraddizione con i principi fondamentali della meccanica newtoniana. Stranamente, tale teoria intuitiva ricorda una teoria della meccanica ritenuta valida da numerosi filosofi nei tre secoli che hanno preceduto Newton.

In una serie di studi compiuti da me e dai miei colleghi alla Johns Hopkins University, è stato chiesto a studenti universitari e delle scuole superiori di prevedere come un oggetto in moto si sarebbe comportato in diverse situazioni. Un esperimento riguardava il moto di un corpo inizialmente impegnato in un percorso circolare, e poi lasciato libero bruscamente. Tale moto ha luogo, per esempio, quando un atleta lancia il disco. Alfonso Caramazza, Bert F. Green Jr. e io stesso abbiamo chiesto ad alcuni studenti universitari di considerare una situazione in cui un uomo lega una palla all'estremità di una corda e la fa ruotare sopra la testa. Che cosa accadrebbe, abbiamo chiesto, se la corda dovesse rompersi? Abbiamo dato agli studenti un diagramma che illustrava il percorso circolare della palla, e abbiamo chiesto loro di tracciare la traiettoria che essa avrebbe seguito dopo la rottura della corda, senza tener conto della resistenza dell'aria. Abbiamo chiesto di mostrare il solo moto orizzontale anche senza considerare la caduta.

Un principio fondamentale della meccanica newtoniana stabilisce che un oggetto si muove lungo una linea retta, a meno che non agiscano su di esso forze esterne. Di conseguenza, la risposta corretta al problema è che la palla si muove in linea retta lungo la tangente alla circonferenza dal momento in cui la corda si rompe. All'incirca un terzo dei 50 studenti a cui abbiamo posto il problema ha, tuttavia, tracciato una traiettoria curva per descrivere il moto della palla. Come dire che essi credono che il moto curvilineo della palla persista per qualche tempo anche dopo che essa viene lasciata libera dal vincolo imposto dalla corda.

Le esperienze riferite da alcuni dei nostri soggetti ci hanno indotto a sospettare che le idee intuitive sul moto hanno un ruolo non solo nel pensiero della gente che si riferisce a situazioni ipotetiche, ma anche nell'interazione con oggetti reali. Per esempio, in uno dei nostri studi, un soggetto, ci ha detto che la prima volta che aveva lanciato un sasso con una fionda aveva rotto un vetro perché, dopo che la pietra era stata lasciata libera, essa non aveva seguito una traiettoria curva come era nelle sue previsioni. Per studiare sistematicamente le relazioni fra convincimento e azione, la mia allieva Deborah Kohl e io abbiamo proposto una situazione semplice, in cui le idee intuitive sul moto circolare potrebbero influire sul modo in cui i soggetti eseguono un compito.

Abbiamo chiesto a un gruppo di studenti di spingere un piccolo oggetto su un tavolo, in modo che esso attraversasse un segmento di corona circolare, di novanta gradi, disegnatovi sopra. L'oggetto era un dischetto con un cuscinetto a sfera che gli permetteva di rotolare sul tavolo. Il segmento di corona circolare era racchiuso fra due archi circolari concentrici e fra due segmenti rettilinei che, se prolungati, si sarebbero incontrati nel centro comune degli archi. Abbiamo detto ai nostri soggetti di incominciare tenendo il dischetto vicino al bordo del tavolo. Essi dovevano spingerlo fino a uno dei segmenti rettilinei, al confine con quello circolare, e lasciarlo libero; il dischetto doveva quindi rotolare fino all'altro bordo rettilineo, senza toccare nessuno degli archi.

Molti compresero che il dischetto poteva attraversare il segmento circolare in linea retta, e cercarono di spingerlo lungo una traiettoria rettilinea. Un quarto dei soggetti, tuttavia, mosse il dischetto lungo una traiettoria curva prima di lanciarlo, nella convinzione che avrebbe continuato così dopo essere stato lasciato libero, in modo da seguire l'arco del segmento circolare. La maggior parte delle persone che tentarono questa seconda strategia rimase visibilmente sorpresa nel vedere che il dischetto non seguiva la traiettoria curva!

È stato relativamente semplice inventare un esperimento in cui le convinzioni intuitive sul moto di corpi lasciati cadere potevano influenzare le azioni di una persona. La mia allieva Allyson Washburn e io abbiamo chiesto ad alcuni studenti universitari di attraversare una stanza e, mentre si muovevano, di lasciar cadere una palla da golf in modo da colpire un bersaglio segnato sul pavimento. Quasi la metà cercò di lasciare la palla quando era direttamente sopra il bersaglio, credendo che essa sarebbe caduta verticalmente e un soggetto sorpassò il bersaglio deliberatamente prima di lasciare la palla, pensando che si sarebbe mossa all'indietro nel cadere.

L'incapacità di comprendere in che modo il moto iniziale di un oggetto trasportato influisca sul suo percorso, una volta lasciato libero, o lanciato, può originare difficoltà anche in altri casi. Nel gioco della pallacanestro, i principianti tirano sistematicamente il pallone troppo forte quando sono in movimento verso il canestro, sbattendolo spesso contro il tabellone. Nel football americano, i *quarterback* dilettanti (*quarterback* è il regista in questo gioco) trovano spesso difficile effettuare un passaggio in corsa. In un racconto sul «Sun» di Baltimora si descriveva come un giocoliere, che effettuava i suoi esercizi in equilibrio su un monociclo, tentasse di insegnare i trucchi del mestiere a un'allieva. Il giocoliere le aveva consigliato di gettare la palla verso l'alto, verticalmente, mentre pedalava sul monociclo, ma essa aveva messo in dubbio la giustezza del consiglio, perché secondo lei la palla sarebbe caduta alle sue spalle, mentre il monociclo proseguiva in avanti.

Come mai la gente sbaglia così spesso nel prevedere il percorso di un oggetto in moto, sia risolvendo problemi a tavolino, sia nella pratica? Numerosi studi recenti, inclusi quelli descritti, mostrano che gli errori non sono casuali, ma sistematici. Essi hanno origine da una teoria del moto, generale e coerente, che guida l'azione in modo adeguato in molte circostanze, ma si discosta nondimeno dai principi della meccanica newtoniana. Sono quindi questi errori, incorporati in una teoria fisica intuitiva, che danno origine a una valutazione sbagliata sul moto. Tale teoria ha una sorprendente somiglianza con quella pre-newtoniana dell'impeto.

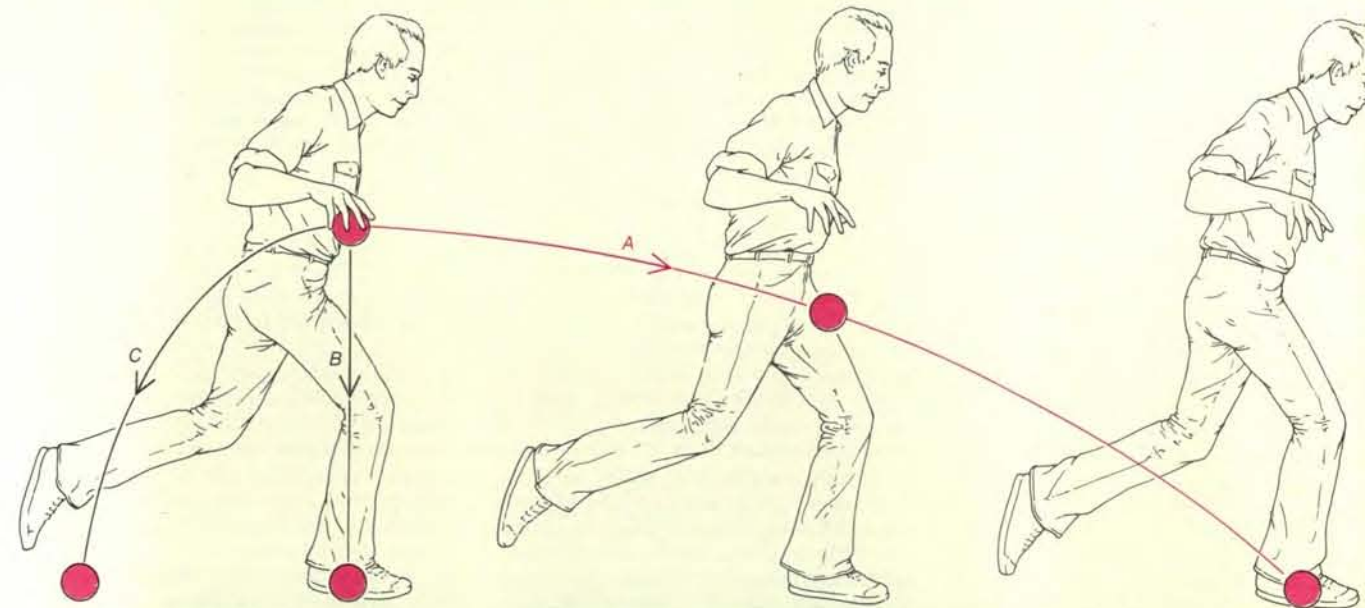
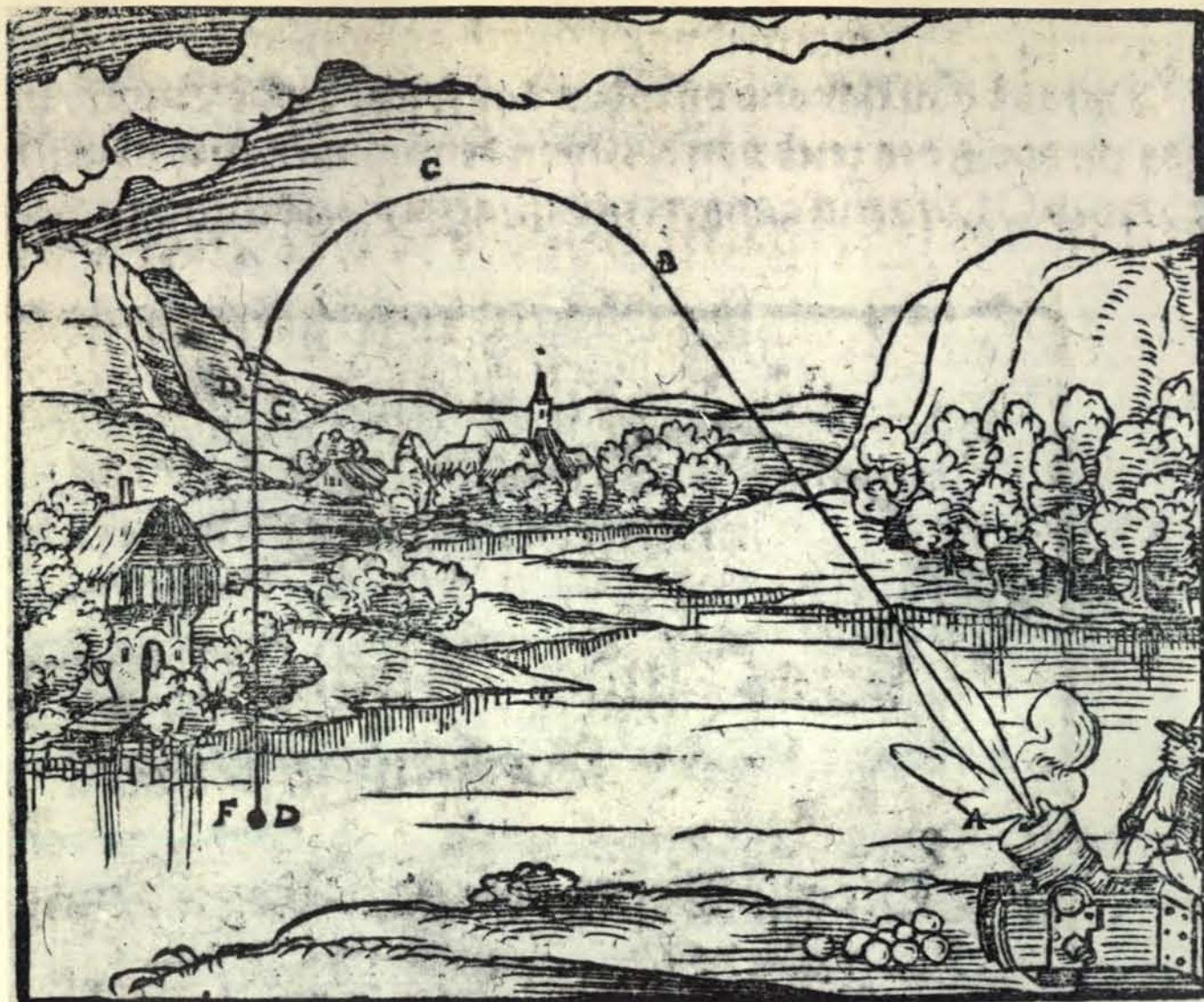
La teoria dell'impeto era una correzione medioevale delle concezioni del moto di Aristotele; essa fu discussa dallo studioso alessandrino Giovanni Filopono, nel VI secolo, e fu sviluppata in modo più completo dal filosofo francese Giovanni Buridano e da altri nel XIV secolo. Buridano ci offre, di esso, una breve enunciazione:

«Il motore, muovendo il mobile, gli imprime un impeto e una certa virtù motrice di quel mobile nella direzione nella quale il motore lo muoveva, sia verso l'alto sia verso il basso, sia lateralmente sia in cerchio... E da quell'impeto è mosso il sasso dopo che il motore ha cessato di muovere» (da Clagett M., *La scienza della meccanica nel Medioevo,* pagina 570, Feltrinelli, Milano, 1972).

La teoria medioevale dell'impeto condivideva con la fisica aristotelica l'idea che il moto dovesse avere una causa. Si pensava che l'azione continua di una forza fosse necessaria a mantenere un oggetto in moto, una concezione esplicitamente negata dalla meccanica newtoniana. Nella fisica aristotelica, la forza responsabile del moto era necessariamente esterna all'oggetto in movimento; questo requisito rendeva piuttosto difficile spiegare il moto di un proiettile, perché una volta che è stato lanciato è difficile trovare per esso una qualsivoglia sorgente di forza esterna. Nella teoria dell'impeto la difficoltà viene aggirata ammettendo che il moto di un proiettile sia conservato da una forza interna all'oggetto, detta impeto, e acquisita dal corpo quando viene messo in moto.

Per spiegare l'osservazione che la maggior parte degli oggetti in movimento giunge alla fine a fermarsi, i teorici avevano anche ammesso che l'impeto andasse gradualmente dissipandosi. Secondo costoro, una palla che rotola su un piano, per esempio, giunge gradualmente a fermarsi perché l'impeto acquisito quando era stata messa in moto diminuisce costantemente fino a esaurirsi del tutto. Alcuni sostenevano che l'impeto si dissipasse spontaneamente, mentre per altri (e tra questi Buridano) la progressiva diminuzione dell'impeto andava attribuita a influenze esterne, come la resistenza dell'aria.

Il fatto che l'impeto si consumasse era una considerazione importante nell'analisi medioevale del moto dei proiettili. Il filosofo arabo Avicenna, nell'XI secolo, sosteneva che un proiettile si muove in linea retta nella direzione in cui viene lanciato fin quando l'impeto iniziale non si esaurisce. Il proiettile, a questo punto, si ferma, e la sua naturale gravità dà origine a un impeto verso il basso che lo fa cadere verticalmente al suolo. Quindi, la traiettoria di un oggetto lanciato orizzontalmente ha la forma di una L rovesciata: l'oggetto si sposta orizzontalmente, si ferma e quindi cade verticalmente.

Una teoria del moto dei proiettili che dà una descrizione in qualche modo più accurata della loro traiettoria fu proposta nel XIV secolo dal filosofo scolastico Alberto di Sassonia. Nella sua spiegazione il moto del proiettile ha tre fasi: nella prima, l'impeto del proiettile sopravanza la sua gravità naturale, e il proiettile si muove lungo una linea retta nella direzione voluta. Tuttavia, l'impeto gradualmente si esaurisce e, a un certo punto, il peso fa deviare il proiettile verso il basso rispetto alla traiettoria iniziale. Così, nella seconda fase, il proiettile si muove sotto l'influenza sia dell'impeto originario, sia

Una palla lasciata cadere da una persona in corsa continua a muoversi in avanti alla stessa velocità del corridore. Il moto in avanti si combina con un moto uniformemente accelerato verso il basso e dà una traiettoria parabolica *(A)*. Le convinzioni intuitive sul moto degli oggetti non sempre corrispondono alla realtà fisica. È stato chiesto a studenti universitari dove toccherebbe il suolo una palla lasciata libera da una persona che cammina. Solo il 45 per cento sapeva che la palla avanzava durante la caduta. Il 49 per cento pensava che la palla cadesse verticalmente, per toccare il suolo sotto il punto in cui era stata lasciata *(B)*; il 6 per cento pensava che la palla si muovesse all'indietro *(C)*.

Secondo la teoria dell'impeto, largamente accettata dal XIV al XVI secolo, quando un oggetto è messo in moto, acquista una forza interna, detta impeto, che tende a conservarlo in moto. La teoria è contraria ai principi fondamentali della meccanica newtoniana, secondo i quali non si richiede alcuna forza per mantenere un oggetto in moto lungo una linea retta a velocità costante. Nondimeno, molta gente ha ancora una comprensione intuitiva del moto dei corpi, che ricorda la teoria dell'impeto. Questo disegno, riprodotto dall'edizione del 1582 del *Bawkunst Oder Architectur aller fürnemsten* di Walther Hermann Ryff, illustra una versione di tale teoria che divide la traiettoria di un proiettile in tre fasi. Nella prima, l'impeto fornito dal cannone predomina sul peso del proiettile e quest'ultimo si muove in linea retta. L'impeto si dissipa gradualmente e il peso fa incurvare verso il basso la traiettoria del proiettile. Nella seconda, il proiettile si muove sotto l'influenza dell'impeto del cannone e dell'impeto dovuto al suo peso. Nella terza, l'impeto del cannone si è esaurito, e il proiettile cade verticalmente.

dell'impeto dovuto al suo peso. Nella terza fase l'impeto originario si è esaurito e il proiettile cade verticalmente.

Per quanto riguarda il moto di un proiettile lanciato verso l'alto, l'astronomo greco Ipparco diede una spiegazione analoga, nel II secolo a.C. Nella sua concezione, l'impeto verso l'alto del proiettile predomina in un primo tempo sul peso, e il proiettile si muove verso l'alto gradatamente mentre l'impeto diminuisce e il moto rallenta. A un certo punto diventa così debole che il peso del proiettile prevale, e ha inizio la caduta. Mentre l'impeto verso l'alto continua a dissiparsi, il proiettile cade con velocità crescente. Nei suoi primi lavori, Galileo adottò la teoria di Ipparco per discutere l'accelerazione di un oggetto in caduta.

È importante notare che i teorici medioevali dell'impeto non si limitarono a trattare il moto rettilineo: secondo la teoria dell'impeto, un oggetto in moto su una traiettoria circolare acquista un impeto circolare, che agisce in modo da conservare tale tipo di moto. Il concetto di impeto circolare veniva invocato per spiegare fenomeni come il persistere della rotazione di una ruota dopo la sua messa in moto e la presunta rotazione delle sfere celesti, in cui si pensava fossero infisse le stelle e i pianeti. Leonardo da Vinci applicò l'idea di impeto circolare per descrivere il moto di un oggetto in condizioni simili a quelle che abbiamo citato per la palla e la corda: «Ogni cosa mossa con furia seguiterà per l'aria la linia del movimento del suo motore. Se quello che muove la cosa in circulo, s'ella fia lasciata in quel moto, il moto suo fia curvo» (Cod. A, 81 V, citato da Leonardo da Vinci, *Scritti scelti*, a cura di Anna Maria Brizio, pagina 264, UTET, Torino, 1966).

La teoria medioevale dell'impeto è incompatibile con la meccanica newtoniana per numerose importanti ragioni. Secondo la prima legge del moto di Newton, come non è richiesta alcuna forza per mantenere un oggetto in quiete, così non è richiesta alcuna forza per mantenerlo in moto a velocità costante (o meglio, a velocità costante lungo una linea retta). Infatti, nella meccanica newtoniana, uno stato di quiete e uno di velocità rettilinea uniforme sono considerati equivalenti: fra di essi non può essere istituita alcuna distinzione fisica, e ogni oggetto che non muta la propria velocità o direzione può essere considerato in quiete o in moto rettilineo uniforme, a seconda della scelta del sistema di riferimento. Per esempio, è equivalente descrivere un passeggero in un aeroplano che non stia accelerando come in quiete, se ci si riferisce all'aeroplano, o in moto se ci si riferisce alla Terra. Il caffè versato in aereo cade nella tazza esattamente come il caffè versato in una condizione di quiete a terra; non finisce verso la coda dell'aereo come dovrebbe se la teoria dell'impeto fosse corretta.

Secondo la teoria dell'impeto, si deve stabilire una distinzione fondamentale fra uno stato di quiete e uno di moto: un oggetto in moto può avere un impeto, un oggetto in quiete no. Come già detto, nessuna distinzione del genere può essere istituita nella meccanica newtoniana. Gli oggetti in moto, tuttavia, hanno una quantità di moto e il concetto di impeto ricorda, così, per qualche aspetto quello di quantità di moto. Ma vi sono due differenze fondamentali: la prima è che l'impeto era considerato una causa del moto, mentre la quantità di moto è solamente una grandezza impiegata per descrivere il moto. La seconda è che, nella teoria dell'impeto, l'impeto di un oggetto è visto in senso assoluto, mentre nella meccanica newtoniana la quantità di moto, come la velocità, è definita rispetto a un sistema di riferimento.

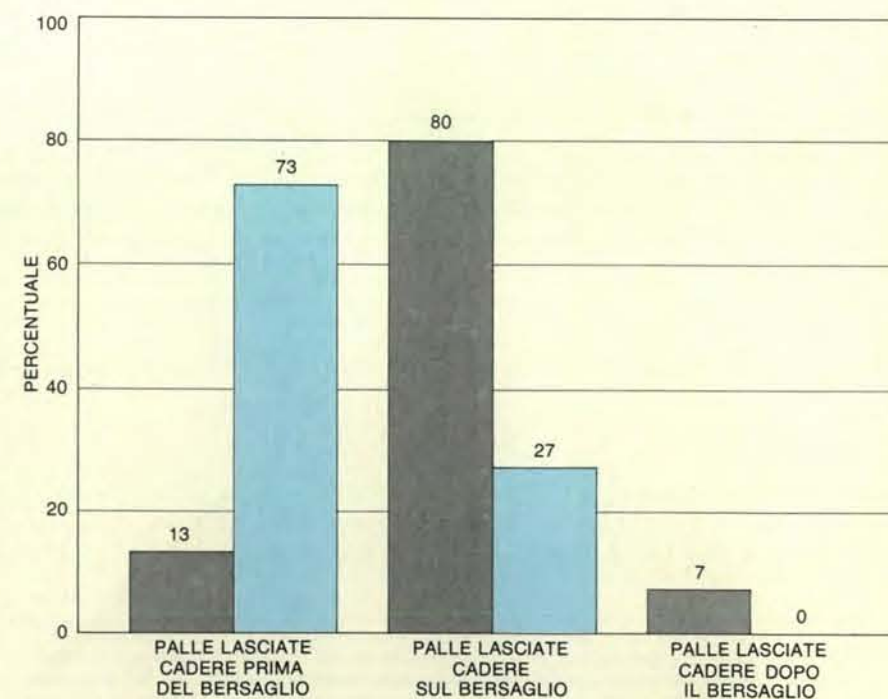
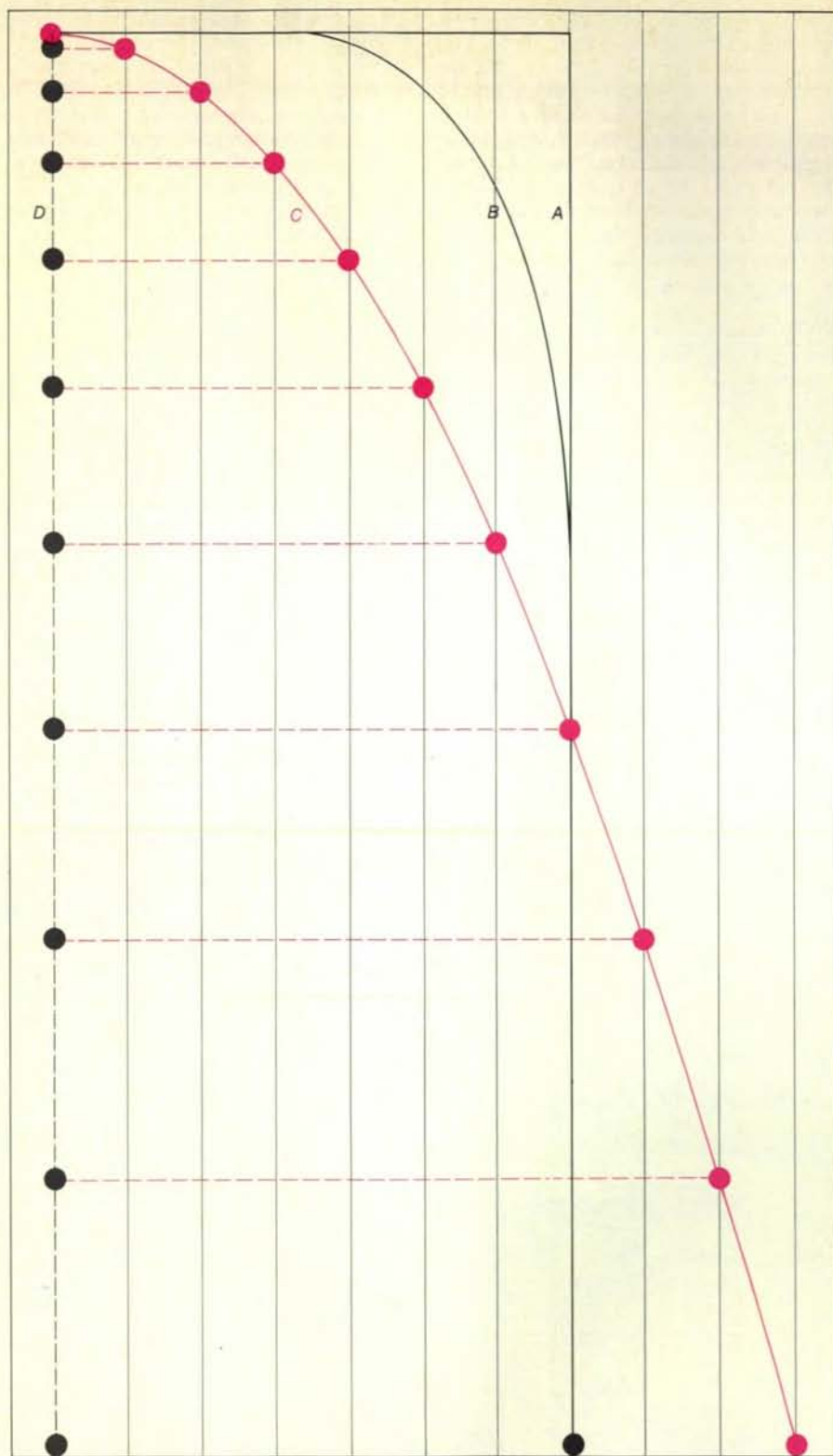
È anche sbagliato supporre che un oggetto in caduta lungo un pendio si fermi o un proiettile cada a causa della perdita di impeto. Nello schema newtoniano gli oggetti in moto si fermano o cominciano a cadere perché intervengono forze esterne a modificare la velocità o la direzione del loro moto. Una palla che rotola su un piano rallenta a causa dell'attrito, una forza che agisce in direzione opposta a quella del moto della palla. Un proiettile sparato orizzontalmente viene accelerato verso il basso dalla forza costante della gravità terrestre. Il moto orizzontale del proiettile è indipendente da quello verticale, sebbene entrambi vengano rallentati dalla resistenza dell'aria. Nel vuoto la velocità orizzontale rimarrebbe costante dal momento in cui il proiettile viene lanciato a quello in cui colpisce il suolo. Per un proiettile sparato verso l'alto, l'azione continua della gravità riduce costantemente la velocità diretta verso l'alto da valori positivi a zero, e quindi da zero a valori negativi, progressivamente più grandi; in altre parole, dopo che il proiettile raggiunge il suo massimo, esso accelera verso il basso.

L'effetto della gravità sul moto verticale del proiettile non è ritardato da alcuna componente orizzontale del moto; il proiettile comincia a deviare verso il basso fin dall'istante in cui viene lanciato. Così, i teorici dell'impeto avevano torto anche nel sostenere che un proiettile si muove lungo una linea retta per qualche tempo dopo essere stato lanciato.

Infine, la meccanica newtoniana e la teoria dell'impeto differiscono per il modo di trattare il moto circolare. Nella teoria dell'impeto il moto circolare è visto come non diverso, fondamentalmente, dal moto rettilineo: entrambe le forme di moto vengono generate impartendo un impeto appropriato a un oggetto. Nella meccanica newtoniana vi è una chiara distinzione fra il moto rettilineo e il moto circolare. Un oggetto si muove di moto rettilineo uniforme quando su di esso non agisce alcuna forza. Affinché un oggetto si muova lungo una circonferenza è, invece, necessario che su di esso agisca una forza esterna che tende a farlo deflettere da un moto rettilineo. Su una palla fissata a un estremo di una corda e fatta muovere in cerchio, agisce la tensione della corda stessa, che trattiene la palla impedendole di allontanarsi a velocità costante lungo una retta tangente alla circonferenza.

Ciascuno degli aspetti importanti della teoria dell'impeto che abbiamo menzionato ha il suo corrispettivo nelle credenze sul moto ancora oggi tanto diffuse. Le idee sul moto della maggior parte delle persone che non hanno ricevuto una specifica istruzione in fisica e quelle di molti che hanno seguito almeno un corso in fisica sono, invero, molto più vicine alle spiegazioni date dalla teoria dell'impeto che a quelle date dalla meccanica newtoniana. Queste sono le conclusioni dei nostri studi alla Johns Hopkins e di esperimenti realizzati da John J. Clement dell'Università del Massachusetts ad Amherst, da Audrey B. Champagne e Leopold E. Klopfer dell'Università di Pittsburgh e da Laurence Viennot dell'Università di Parigi.

In uno studio realizzato dalla Kohl, dalla Washburn e da me, studenti delle scuole superiori e universitari hanno tentato di risolvere parecchi problemi sul moto, e quindi di spiegare le loro risposte. In molti casi essi enunciarono quasi esplicitamente la teoria dell'impeto. Ad esempio, a uno studente che usava il termine «quantità di moto» per spiegare le sue risposte, fu chiesto di darne una definizione. Rispose: «È un qualcosa che trasporta [un oggetto] dopo che la forza su questo ha terminato di agire. È una forza di moto. Qualcosa che mantiene l'oggetto in moto». Un altro studente descrisse la quantità di moto di una palla così: «[È] una forza applicata e introdotta nella palla in modo che essa, ora che è in moto, abbia una certa quantità di forza. L'oggetto in moto ha la forza della quantità di moto e, poiché nulla si oppone a tale forza, esso continuerà nel suo moto fin quando non sarà fermato». In queste due affermazioni la tesi che il moto venga conservato da una forza interna è chiara.

Anche l'idea che gli oggetti in moto rallentino e si fermino a causa della perdita dell'impeto ha trovato un'espressione nelle spiegazioni dei soggetti esaminati. Uno in particolare, a cui era stato chiesto di spiegare come mai una palla che rotola su un piano a un certo punto si fermi, diede la seguente risposta: «Capisco che l'attrito e la resistenza dell'aria influisca-

Le convinzioni intuitive sul moto possono influenzare i modi in cui gli individui interagiscono con gli oggetti in movimento. È stato chiesto ad alcuni studenti universitari di attraversare una stanza e di lasciar cadere, camminando, una palla da golf su un bersaglio segnato sul pavimento. I risultati, riassunti dagli istogrammi, mostrano che una preparazione in fisica influenza le strategie che vengono adottate per colpire il bersaglio. I soggetti che avevano seguito almeno un corso di fisica *(in colore)* erano più propensi a lasciar cadere la palla prima di raggiungere il bersaglio di quelli che non avevano alcuna istruzione del genere *(in grigio)*. Più di un quarto dei soggetti con un'istruzione in fisica pensò, tuttavia, erroneamente che la palla doveva essere lasciata cadere quando era sopra il bersaglio. Le conclusioni mostrano che gli studenti spesso considerano la teoria intuitiva dell'impeto come uno schema per interpretare il materiale fornito loro nei corsi di meccanica newtoniana e deformano così le nuove nozioni per farle collimare con essa.

Le convinzioni espresse sul moto dei proiettili da persone sottoposte a un esperimento ricordano la teoria medioevale dell'impeto. È stato chiesto ad alcuni studenti di disegnare la traiettoria di un proiettile lanciato orizzontalmente, trascurando la resistenza dell'aria. Il 5 per cento ha risposto che il proiettile si muove prima orizzontalmente e poi verticalmente verso il basso *(A)*: una posizione simile a quella del filosofo arabo dell'XI secolo Avicenna. Il 35 per cento credeva che il proiettile si muovesse prima lungo una retta orizzontale, quindi cominciasse a incurvare la sua traiettoria verso il basso e, infine, cadesse verticalmente *(B)*, in accordo con la teoria a tre fasi sul moto dei proiettili del filosofo scolastico del XIV secolo Alberto di Sassonia. Il 28 per cento ha disegnato la traiettoria corretta, che è una parabola *(C)*; la maggior parte del restante 32 per cento ha indicato che il proiettile comincia subito a cadere, ma pensava erroneamente che, alla fine, sarebbe caduto in verticale. La parabola rappresenta la combinazione di un moto orizzontale uniforme e di un moto verticale accelerato. Il tratto verticale percorso da un proiettile in caduta durante ciascun intervallo di tempo è lo stesso percorso da un oggetto in quiete, lasciato cadere *(D)*. Così, un proiettile sparato orizzontalmente e uno lasciato cadere dalla stessa altezza simultaneamente colpiscono il terreno nello stesso istante.

no negativamente sulla velocità della palla, ma non so come. Non sono sicuro se essi selezionino o assorbano parte della forza presente nella palla. In altre parole, se la palla debba impiegare o consumare della forza, e quindi perderla, per vincere la resistenza dell'aria o l'attrito». L'intervistato ammette, anche solo come ipotesi, che l'attrito e la resistenza dell'aria esauriscano l'impeto della palla.

L'idea intuitiva che un oggetto lasciato libero da un moto circolare segua per qualche tempo un percorso curvilineo, sembra aver origine dall'idea che tale oggetto abbia un impeto circolare. Un soggetto che aveva disegnato una traiettoria curva per la palla fissata all'estremità di una corda, che abbiamo menzionato prima, spiegò che la palla seguiva la traiettoria curva «a causa della quantità di moto direzionale. È presente una forza circolare e, [dopo che la corda si rompe, la palla] seguirà la traiettoria disegnata fino a quando non rimarrà priva della forza nel suo interno, forza che è stata creata per rotazione». In altre parole, la palla fatta ruotare acquista un impeto circolare che le fa seguire una traiettoria circolare dopo che la corda si è spezzata. Ma tale impeto si consuma e la traiettoria gradualmente diventa rettilinea.

I nostri soggetti hanno ricapitolato la teoria medioevale dell'impeto anche nelle loro credenze sul moto dei proiettili. Quando è stato chiesto loro di spiegare come mai i proiettili raggiungano un'altezza massima e quindi cadano, essi hanno assunto spesso la posizione inizialmente proposta da Ipparco. Per esempio, un soggetto spiegò il moto verticale di una palla di cannone nel modo seguente: «Mentre la palla sale, la forza impressa dal cannone si dissipa e quella di gravità prende il sopravvento. Così, la palla rallenta. Nel punto più alto della traiettoria, la forza impressa dal cannone e quella di gravità sono circa uguali. Questo è il punto di inversione, nel quale la forza del cannone diminuisce, mentre la gravità diviene la forza predominante. Mentre la palla completa l'arco e comincia a scendere, la forza di gravità ha la meglio». Clement ha trovato che alcuni studenti danno spiegazioni analoghe per il moto di una monetina lanciata in aria.

Alcuni soggetti che dettero questo genere di spiegazione affermarono che un proiettile, cadendo, accelera perché continua a perdere l'impeto originario verso l'alto. Forse è bene ricordare, però, che, nella meccanica newtoniana, non vi è né impeto verso l'alto né una forza verso l'alto; vi è solo la componente verso l'alto della velocità del proiettile, che si riduce costantemente per l'azione della gravità. Il culmine nella traiettoria del proiettile non è un punto in cui le forze verso l'alto e verso il basso sono uguali, come succede nella teoria dell'impeto.

In uno dei problemi che abbiamo posto agli studenti sul moto di un proiettile, abbiamo chiesto di supporre che una palla di metallo venisse spinta ad alta velocità verso la sommità di un dirupo, superandone quindi il bordo. Abbiamo poi richiesto di

«Mobile grave ad hoc ut violenter sursum moveri possit necessaria est virtus impellens maior gravitate resistenti, alioquin gravitas resistens non posset vinci, nec consequenter grave sursum ferri; fertur ergo sursum mobile dum motiva virtus impressa maior est gravitate resistente; cum autem ista virtus, ut demonstratum est continue debilitetur deveniet tandem adeo diminuta ut mobilis gravitatem non amplius vincet, et tunc non ulterius impellet mobile...
«...sed amplius suo more decrescente virtute impressa incipit predominari gravitas mobilis quare et mobile descendere aggreditur, sed quia in principio talis descensus superest adhuc multa virtus sursum impellens (licet non amplius maior sit gravitate mobilis) quae est levitas hinc fit ut gravitas propria mobilis ab hac levitate minuatur, et consequenter motus tardior sit in principio, et rursus quia extrinseca illa virtus amplius debilitatur, minorem gravitas mobilis resistentiam habens augetur et mobile velocius adhuc movetur; et hanc veram existimo causam accelerationis motus...»

«Perché un mobile grave possa muoversi di moto violento verso l'alto è necessaria una virtù impellente maggiore della gravità resistente, giacché in caso contrario la gravità che resiste non potrebbe esser vinta né, di conseguenza, il grave potrebbe muoversi verso l'alto. Il mobile si muove dunque verso l'alto finché la virtù impressa è maggiore della gravità che oppone resistenza; poiché però questa virtù, come è stato dimostrato, si indebolisce in modo continuo, diverrà infine tanto piccola da non vincere più la gravità del mobile, e quindi da non spingere più il mobile verso l'alto...
«...ma decrescendo ancor più, com'è suo costume, la virtù impressa, comincia a predominare la gravità del mobile, e perciò il mobile comincia a discendere; ma poiché nel principio di tale discesa rimane ancora molta virtù che spinge verso l'alto (pur non essendo più maggiore della gravità del mobile), che è la leggerezza, ne consegue che la gravità propria del mobile è diminuita da questa leggerezza, e di conseguenza che il moto è più lento in principio; e di nuovo, poiché quella virtù estrinseca si indebolisce ancor più, la gravità del mobile, avendo una minore resistenza, aumenta, e il mobile si muove ancor più veloce; e questa stimo sia la vera causa dell'accelerazione del moto» *(traduzione di Libero Sosio).*

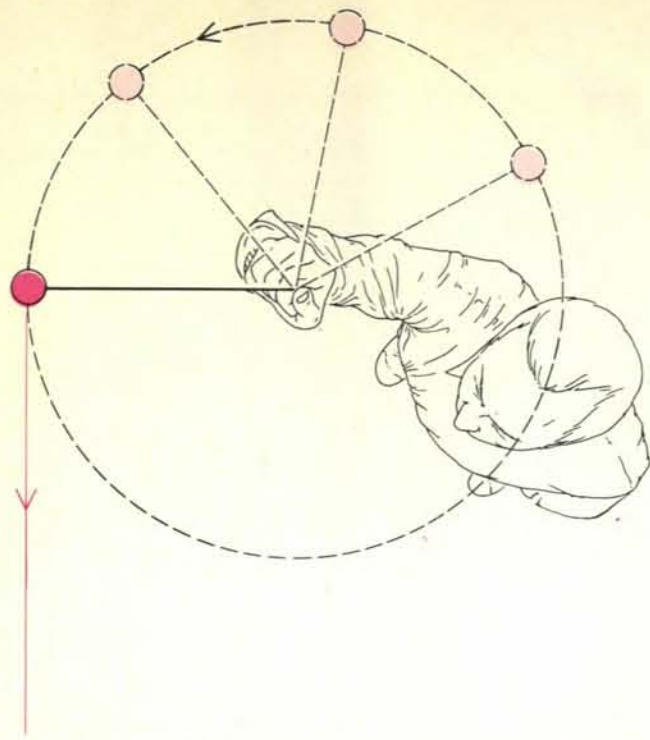
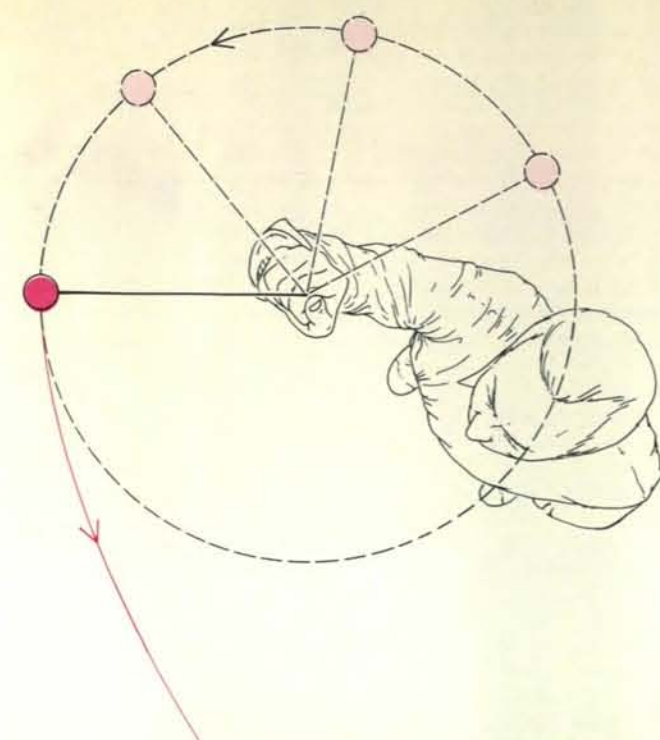
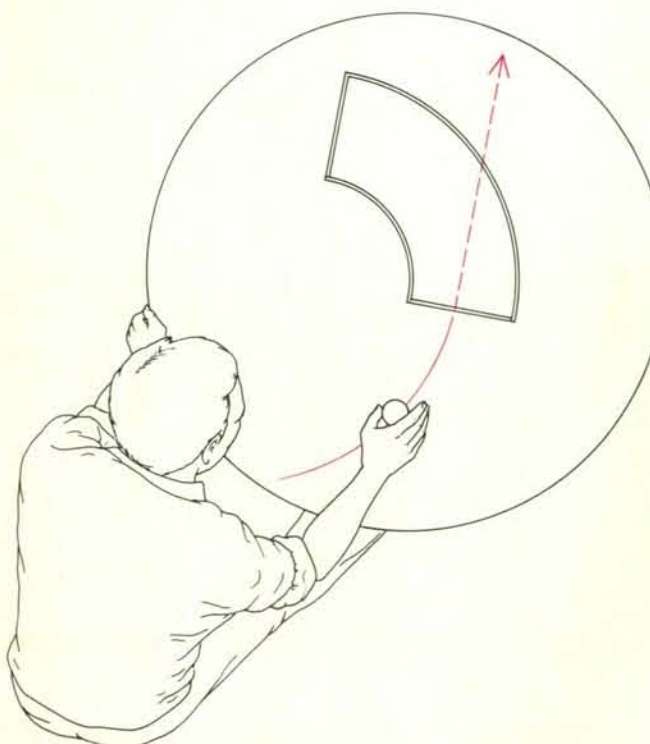
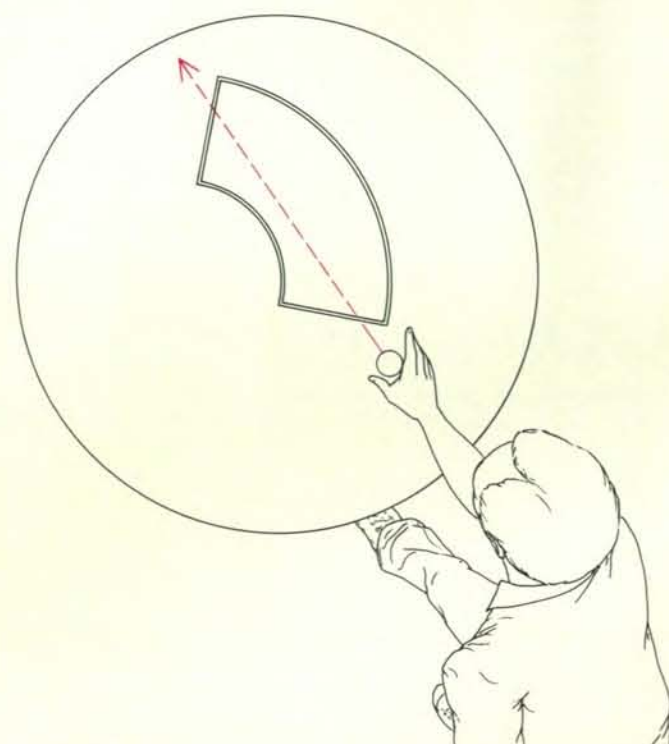
Il manoscritto del *De motu* di Galileo, un saggio scritto intorno al 1590 e conservato alla Biblioteca nazionale centrale di Firenze, dimostra come lo scienziato pisano, nei suoi primi lavori, avesse accettato la teoria dell'impeto. Nel brano, egli applica il concetto di impeto per spiegare il moto di un proiettile lanciato verso l'alto. L'autore e i suoi colleghi della Johns Hopkins University e John J. Clement dell'Università del Massachusetts ad Amherst hanno trovato che i soggetti ai quali veniva chiesto di spiegare la salita e la caduta di un proiettile davano spesso spiegazioni simili.

tracciare la traiettoria della palla nella sua caduta, ignorando la resistenza dell'aria. La risposta corretta è che la velocità orizzontale costante della palla e la sua velocità verticale, in costante aumento, si combinano dando un arco di parabola.

Più di un terzo dei soggetti a cui era stato presentato il problema sostenne che la palla si muove in linea retta, orizzontalmente, per una certa distanza al di là del bordo della rupe. Successivamente segue una traiettoria curva e infine cade verticalmente. Un soggetto diede la seguente spiegazione: «Quando [la palla] lascia il dirupo, la forza orizzontale è più intensa di quella verso il basso. Quando diviene inferiore, il corpo comincia a cadere». Il cinque per cento dei nostri intervistati si rifece alla teoria di Avicenna: a loro giudizio la palla si spostava prima orizzontalmente, e poi verticalmente. Uno di costoro difese il suo ragionamento dicendo che «la gravità non agisce fino a quando il corpo non smette di muoversi».

Anche la convinzione che un oggetto trasportato cada verticalmente al suolo quando viene lasciato libero deriva dalla teoria intuitiva dell'impeto. Chi la sostiene spiega che un oggetto acquista impeto quando viene spinto o lanciato, ma non quando viene trasportato passivamente. Quindi, quando un oggetto trasportato viene lasciato cadere, non vi è una forza che lo faccia muovere in avanti, e così esso cade verticalmente. Questa convinzione intuitiva viene mantenuta anche in situazioni in cui il mezzo di trasporto si muove ad alta velocità. In un esperimento abbiamo chiesto agli studenti di tracciare la traiettoria di una bomba lasciata cadere da un aeroplano in volo, ignorando gli effetti della resistenza dell'aria. Più di un terzo disse che la bomba sarebbe caduta verticalmente.

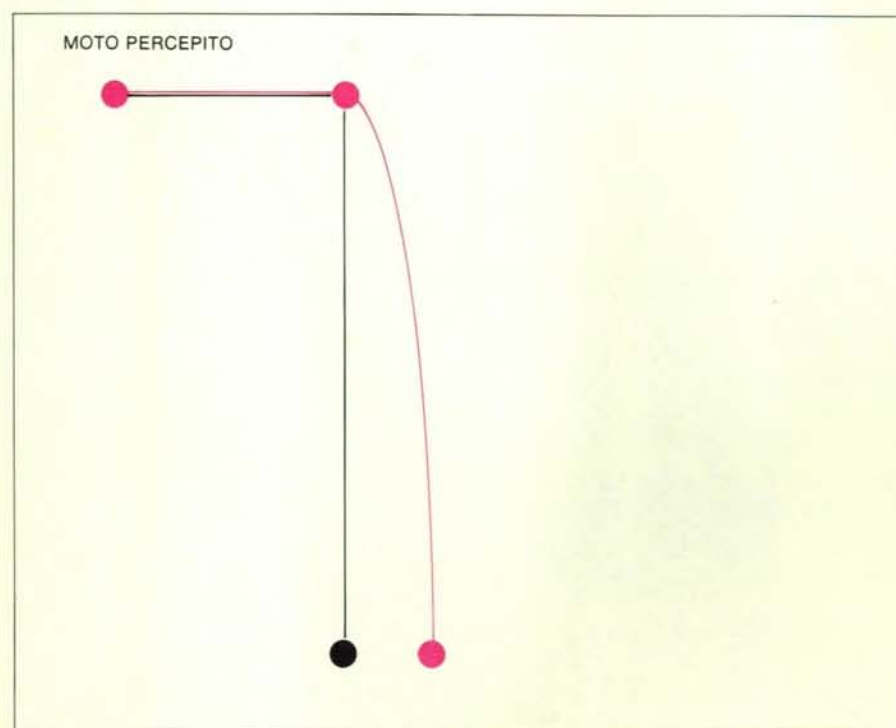
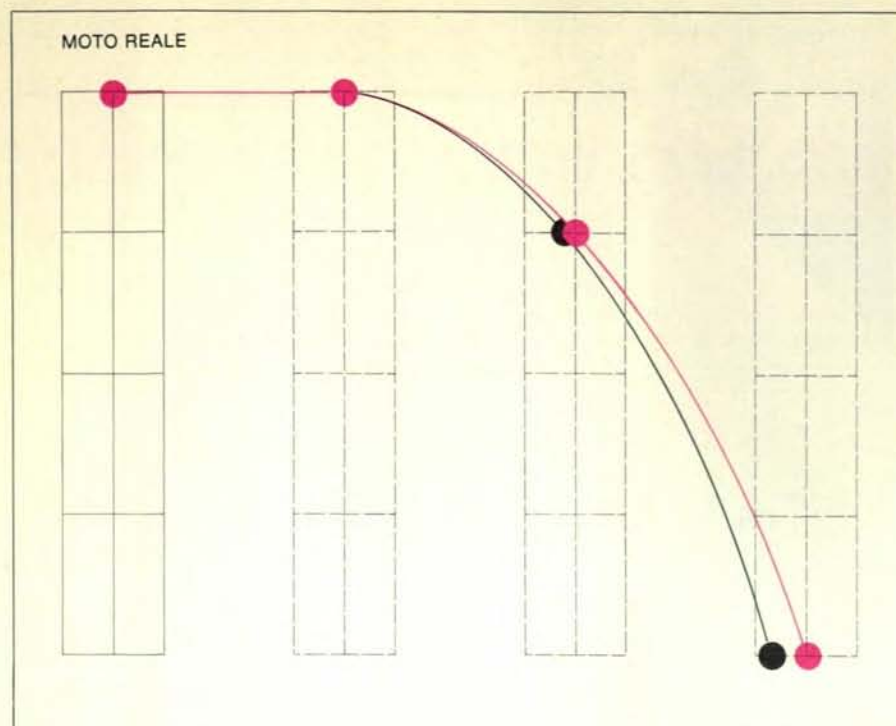
La sorprendente somiglianza fra le opinioni dei filosofi medioevali e quelle dei nostri intervistati fa pensare che la teoria dell'impeto sia il risultato naturale delle nostre esperienze con i moti sulla Terra. È facile vedere come l'esperienza potrebbe portare al principio fondamentale che gli oggetti sono tenuti in moto da una forza interna che viene gradualmente consumata: nella maggior parte dei casi, infatti, un oggetto messo in moto continua a muoversi per un certo tempo, ma alla fine si ferma. È difficile capire, invece, come mai la gente sviluppi opinioni sbagliate sulle traiettorie di oggetti in moto, opinioni che contraddicono apertamente l'esperienza quotidiana. Perché, ad esempio, la gente giunge a credere che oggetti trasportati cadono verticalmente quando vengono lasciati liberi?

Una possibile risposta è che, in certe circostanze, il moto degli oggetti viene percepito sistematicamente in modo erroneo. La Washburn, Linda Felch e io abbiamo ipotizzato che gli oggetti lasciati cadere da un mezzo in moto vengano percepiti come se cadessero verticalmente. La percezione erronea può essere all'origine delle opinioni, anch'esse erronee, sul

Secondo la teoria dell'impeto, quando un corpo si muove lungo un cerchio, gli viene impartito un impeto circolare. La teoria dell'impeto prevede quindi, erroneamente, che una palla fatta ruotare a un capo di una corda continuerà a seguire un percorso curvilineo nel caso che la corda si spezzi. In realtà, a cominciare dall'istante in cui si rompe la corda, la palla si muoverà lungo una linea retta tangente al suo percorso circolare originario. Il 51 per cento degli studenti universitari, a cui è stato chiesto di tracciare la traiettoria della palla dopo la rottura della corda, ha risposto correttamente *(a sinistra)*. Il 30 per cento, pensava, invece, che il moto circolare sarebbe continuato e che la traiettoria sarebbe risultata curva *(a destra)*; il 19 per cento ha dato risposte erronee di altro genere.

Alcune persone che hanno a che fare con oggetti in moto si comportano come se a questi potesse essere impartito un impeto circolare. Nel nostro esperimento, abbiamo disegnato su un tavolo un segmento di corona circolare di 90 gradi e abbiamo dato ad alcuni studenti un dischetto, munito di cuscinetti a sfera, che poteva rotolare sul tavolo. Abbiamo chiesto di spingere il dischetto fino a un estremo del segmento circolare, e quindi di lasciarlo libero; il compito si prefiggeva di far giungere il dischetto all'altro estremo del segmento senza toccare i lati curvi. Il 25 per cento degli studenti ha adottato la strategia illustrata a sinistra: hanno mosso il dischetto lungo un arco, seguendo evidentemente la convinzione erronea che l'oggetto avrebbe continuato a muoversi lungo una traiettoria curvilinea mentre si avvicinava al segmento circolare. La linea tratteggiata, che rappresenta la traiettoria del dischetto dopo che è stato lasciato libero, mostra come questa strategia sia fallita. Il 67 per cento dei soggetti, invece, ha applicato la strategia corretta che è illustrata a destra: essi tiravano il dischetto tenendo presente che avrebbe seguito una traiettoria rettilinea dopo essere stato lasciato libero. Il rimanente 8 per cento ha tentato altri metodi senza successo.

comportamento di oggetti in caduta. Studi sulla percezione del moto hanno mostrato che si può avere un'illusione ottica osservando un oggetto rispetto a un sistema di riferimento in movimento. Il moto dell'oggetto rispetto al sistema di riferimento mobile può essere erroneamente percepito come moto assoluto (cioè come moto relativo a un ambiente stazionario.) Per esempio, se un punto all'interno di un rettangolo rimane fisso mentre il rettangolo si muove verso destra, il punto viene visto come se si muovesse verso sinistra.

Analogamente, quando si osserva un oggetto trasportato che viene lasciato cadere, il corpo che lo trasporta rimane di solito in vista mentre esso cade; il mezzo di trasporto può quindi essere preso come un sistema di riferimento. In situazioni in cui gli effetti della resistenza dell'aria sono trascurabili, l'oggetto lasciato libero cade verticalmente rispetto al mezzo di trasporto. Da qui potrebbe nascere l'illusione che l'oggetto cade verticalmente rispetto al suolo.

Allo scopo di verificare quest'ipotesi abbiamo chiesto ad alcuni soggetti di osservare delle visualizzazioni al calcolatore che simulavano situazioni in cui oggetti trasportati venivano lasciati cadere. Un riquadro rettangolare, che rappresentava il mezzo di trasporto, si muoveva attraverso lo schermo con una velocità moderata e costante da sinistra a destra. Un dischetto che rappresentava l'oggetto trasportato, rimaneva sulla sommità del riquadro, al centro, fino a quando questo non aveva coperto circa un terzo della distanza attraverso lo schermo. Il dischetto tracciava poi uno dei vari percorsi che erano stati studiati per simulare la traiettoria di un oggetto lasciato cadere in assenza o in presenza di resistenza dell'aria. Durante la sua caduta, il riquadro continuava a muoversi con la velocità iniziale. In ciascuna visualizzazione il soggetto doveva prima osservare il moto e quindi tentare di disegnare il percorso seguito dal dischetto.

I nostri risultati indicarono che gli intervistati erano soggetti a una forte illusione ottica quando guardavano le dimostrazioni. Quando il dischetto tracciava il percorso di un corpo in caduta nel vuoto, gli intervistati tracciavano una linea che era generalmente molto più ripida di quella che avevano appena osservato. Come dire che essi percepivano il percorso del dischetto come quasi verticale, con una lievissima curvatura nella direzione del moto del riquadro, nonostante che esso si fosse spostato molto di più verso destra. Quando il dischetto cadeva come un oggetto lievemente influenzato dalla resistenza dell'aria, veniva percepito come se cadesse verticalmente, sebbene si muovesse in realtà, ancora una volta, molto più a destra. L'illusione era presente solo quando veniva mostrato il riquadro in movimento; quando i soggetti osservavano lo stesso moto del dischetto in assenza del sistema di riferimento in movimento, l'illusione non aveva luogo e i loro disegni erano accurati.

In un secondo esperimento i nostri soggetti osservarono sul video la registrazione

Le illusioni ottiche nella percezione del moto possono spiegare alcune concezioni erronee sul moto dei corpi, ancora così diffuse. L'autore e due sue allieve, Allyson Washburn e Linda Felch, hanno verificato la rilevanza di quest'ipotesi nello spiegare l'idea errata che un corpo lasciato libero da un oggetto in moto cada verticalmente. Un riquadro rettangolare, che rappresentava il mezzo di trasporto, si muoveva da sinistra a destra sullo schermo di un calcolatore. Un dischetto, che rappresentava l'oggetto trasportato, rimaneva fisso sulla sommità del riquadro, al centro, per circa un terzo della distanza sullo schermo, e poi cominciava a cadere, muovendosi come un oggetto in caduta nel vuoto *(curva in colore in alto)*, o come un oggetto lievemente influenzato dalla resistenza dell'aria *(curva in nero in alto)*. Dopo aver osservato il moto del dischetto e del riquadro rettangolare, i soggetti sottoposti agli esperimenti hanno tentato di disegnare la traiettoria che il dischetto aveva appena tracciato sullo schermo. Quando il dischetto cadeva come se si trovasse in assenza d'aria, esso veniva generalmente percepito come se, nel corso della caduta, si spostasse solo leggermente in avanti *(curva in colore in basso)*. Quando si muoveva come se fosse influenzato dalla resistenza dell'aria, la traiettoria media percepita era una linea retta verticale *(curva nera in basso)*. Tale illusione ottica non aveva luogo quando i soggetti osservavano gli stessi tipi di moto del dischetto in assenza del riquadro in movimento sullo sfondo.

di una persona che, mentre camminava, lasciava cadere una palla; di nuovo, il compito loro assegnato era quello di tracciare il percorso della palla. Essi erano stati avvisati che gli eventi nella registrazione potevano essere stati truccati, anche se in realtà ciò non era vero. Di conseguenza, era stato detto loro, dovevano tracciare esattamente ciò che vedevano, non quello che pensavano si sarebbe verificato. Dei 18 soggetti che videro la registrazione, solo quattro indicarono che la palla si muoveva in avanti mentre cadeva. Dieci soggetti dissero che la palla cadeva verticalmente, e quattro che si muoveva all'indietro.

Tali risultati non dimostrano che la causa delle convinzioni erronee sul moto di oggetti lasciati cadere sia una percezione erronea, ma suggeriscono che l'ipotesi sia (senz'altro) ragionevole. I miei colleghi e io stiamo studiando la possibilità che le illusioni siano coinvolte nello sviluppo di altre concezioni erronee sul moto.

Come è possibile correggere queste concezioni erronee? La risposta più ovvia è: con una normale istruzione sulla meccanica newtoniana. Ma gli studi di diversi ricercatori fanno pensare che sia difficile modificare le idee intuitive. Sebbene alcuni studenti che scelgono corsi di fisica raggiungano una buona comprensione della meccanica newtoniana, molti ne escono con le teorie dell'impeto ancora intatte. La Kohl, la Washburn e io abbiamo saggiato le conoscenze degli studenti di scuole superiori sul moto prima e dopo che seguissero un corso di fisica, trovando che esso aveva lasciato intatte alcune concezioni erronee. Il corso, in particolare, non riuscì a intaccare il nucleo centrale dell'assunto che l'impeto, acquistato quando un oggetto viene messo in moto, serve a mantenere il moto stesso; l'ottanta per cento degli studenti conservò questa convinzione.

Sembra che gli studenti si basino sulla teoria intuitiva dell'impeto come schema generale per interpretare nuovo materiale fornito dal corso. Di conseguenza, questo materiale può essere alterato e distorto per farlo collimare con i preconcetti intuitivi. Numerosi insegnanti stanno cercando vie per strutturare l'istruzione in fisica in modo da evitare tali tranelli. James Minstrell, un insegnante di scuola superiore a Mercer Island, nello stato di Washington, ha avuto un certo successo con un metodo per mezzo del quale gli studenti sono portati a formulare le loro concezioni intuitive. Le differenze fra queste convinzioni e i principi della meccanica newtoniana possono poi venire evidenziate. Il lavoro di Andrea A. di Sessa, del Massachusetts Institute of Technology, suggerisce che anche l'esperienza con i giochi al calcolatore, in cui gli oggetti si comportano come se si muovessero in un mondo newtoniano senza attriti, può risultare utile.

Quindi lo studio delle teorie intuitive e dei processi con cui esse vengono acquisite e modificate, si dimostra promettente per lo sviluppo di migliori metodi educativi e può aiutarci anche a mettere in luce i fattori che hanno influenzato lo sviluppo storico di teorie scientifiche formali.

L'attività del cervello durante il sonno

La fase REM, cioè il sonno con movimenti oculari rapidi in cui compare l'attività onirica, si associa normalmente a paralisi. Questa si può eliminare negli animali, la cui fase REM può così essere analizzata

di Adrian R. Morrison

Il sonno è comunemente ritenuto una condizione di completo rilassamento e di inattività. Tuttavia, osservando l'attività elettrica del cervello di un essere umano, o di un animale da esperimento, che dorme, risulta evidente che il sonno è una condizione complessa ed è ben lungi dall'essere uno stato di inattività. In effetti, nella fase del sonno in cui compare l'attività onirica, cioè nel sonno REM (dall'inglese *Rapid Eye Movement*, movimento oculare rapido), l'attività elettrica del cervello è molto più simile a quella della veglia che non a quella delle altre fasi del sonno. Nel sonno REM variazioni transitorie di tensione elettrica stanno a indicare che il cervello è in uno stato di risveglio, nonostante sia molto ridotto il contatto sensoriale tra l'animale e il suo ambiente. Ciò che impedisce che il risveglio dei neuroni (cellule nervose) cerebrali durante il sonno REM venga tradotto in movimenti fisici vigorosi è l'assenza di attività nei muscoli dell'animale, la quale si manifesta come una paralisi che dura fino alla fine dell'episodio di sonno REM.

Che cosa accadrebbe se la paralisi che accompagna il sonno REM fosse eliminata e il cervello così risvegliato fosse in grado di attivare i muscoli? La ricerca che si svolge nel mio laboratorio presso l'Università della Pennsylvania e in altri laboratori comincia a dare una risposta a questa domanda. La parte più primitiva del cervello è il tronco cerebrale, situato tra il midollo spinale e la restante parte del cervello. Piccole lesioni eseguite con molta precisione nel gatto, a livello del tronco cerebrale, fanno sì che l'animale mostri le caratteristiche elettriche del sonno REM, pur muovendosi vigorosamente. Esse vengono eseguite nella zona del tronco denominata ponte, e la loro esatta posizione determina in quali muscoli venga eliminata la paralisi.

Quindi la perdita del tono muscolare nel sonno appare direttamente attribuibile al ponte. Altri risultati sperimentali suggeriscono che i neuroni di questa parte del tronco cerebrale influiscono indirettamente anche sulla locomozione. Durante il sonno REM, sembra che una regione del ponte interferisca con l'azione di un «centro della locomozione» che si estende dal ponte a un punto molto più in basso nel tronco cerebrale. Questa interferenza impedisce che tale centro attivi le reti nervose del midollo spinale che sono responsabili del movimento reciproco (alternante) degli arti. Così l'attività del ponte durante il sonno REM inibisce il movimento in due modi. La mia ricerca suggerisce che la coesistenza di una maggiore attività del cervello e di una paralisi muscolare possa verificarsi anche nella veglia e possa essere la causa di alcuni disturbi del sonno.

La transizione dallo stato di veglia e di allerta al sonno comporta molte modificazioni nella fisiologia e nella postura dell'animale; queste alterazioni danno luogo complessivamente a un profondo cambiamento di stato. La mia ricerca è stata eseguita in prevalenza sul gatto e le modificazioni qui descritte si riferiscono in particolare a questo animale, anche se molte di esse sono estendibili ad altre specie di mammiferi. Quando un gatto si addormenta, assume la ben nota posizione raggomitolata e rilassata. I suoi occhi si chiudono e la membrana nittitante (la cosiddetta «terza palpebra») copre parte dell'occhio al di sotto delle palpebre esterne. Il numero delle fibre muscolari in contrazione è ridotto e quindi si nota anche una riduzione del tono muscolare. Nella fase iniziale del sonno scompaiono i movimenti dell'occhio che si possono rilevare nell'animale sveglio e si verifica anche una graduale diminuzione della temperatura del cervello, in genere di una frazione di grado centigrado.

Al di sotto di queste modificazioni, vi sono anche le modificazioni dell'attività dei neuroni del cervello, che hanno luogo quando comincia il sonno. La fonte di informazione più comune riguardo all'attività elettrica di quest'organo è l'elettroencefalogramma (EEG), che consiste nella registrazione dei suoi continui cambiamenti di potenziale elettrico, misurati tramite elettrodi applicati sul cranio. I rapidi cambiamenti di potenziale rappresentano la somma dei cambiamenti di tensione a livello delle membrane esterne di un gran numero di neuroni cerebrali.

L'EEG di un gatto sveglio in allerta mostra onde con un'ampiezza relativamente bassa e una frequenza relativamente alta. Quando il gatto si raggomitola e si addormenta subentra una notevole modificazione: le onde della veglia con bassa ampiezza sono sostituite gradualmente da onde con ampiezza elevata e bassa frequenza. Grazie alla presenza di quest'ultimo aspetto dell'EEG, la fase iniziale leggera del sonno viene spesso indicata come «sonno a onde lente». Quando l'animale entra in questa fase, si verificano anche cambiamenti dell'attività elettrica di aree specifiche delle zone più rostrali o più caudali del cervello, cambiamenti che possono venire registrati tramite elettrodi inseriti in adeguate strutture cerebrali.

Per molti anni si è ritenuto che l'aspetto a onde lente prevalesse lungo tutta la durata del sonno. Nel 1953, tuttavia, Eugene Aserinsky e Nathaniel Kleitman dell'Università di Chicago notarono in soggetti umani intervalli di sonno in cui la registrazione elettroencefalografica riproponeva l'aspetto ad alta frequenza e a bassa ampiezza della veglia. Dopo tale modificazione elettroencefalografica, negli occhi dei soggetti che dormono sono stati osservati movimenti rapidi in varie direzioni, che hanno fatto designare tale fase come sonno REM. Sappiamo ora che periodi REM ricorrono regolarmente durante il sonno, alternandosi a periodi più lunghi di sonno a onde lente. La cadenza e la durata dei periodi REM variano con le specie. Nell'uomo, essi si verificano circa una volta ogni 90 minuti e nel gatto circa una volta ogni 25 minuti. Ciascun episodio dura parecchi minuti.

Svegliando una persona che dormiva immediatamente dopo il termine di un episodio REM, Aserinsky e Kleitman poterono dimostrare che il sonno REM è associato a un'intensa attività onirica. L'iniziale eccitamento che si ebbe per la scoperta del sonno REM era giustificato in gran parte dal fatto di aver identificato il periodo, durante il sonno, in cui si sogna. In seguito, la ricerca ha mostrato che gli aspetti fisiologici del sonno REM sono altrettanto interessanti.

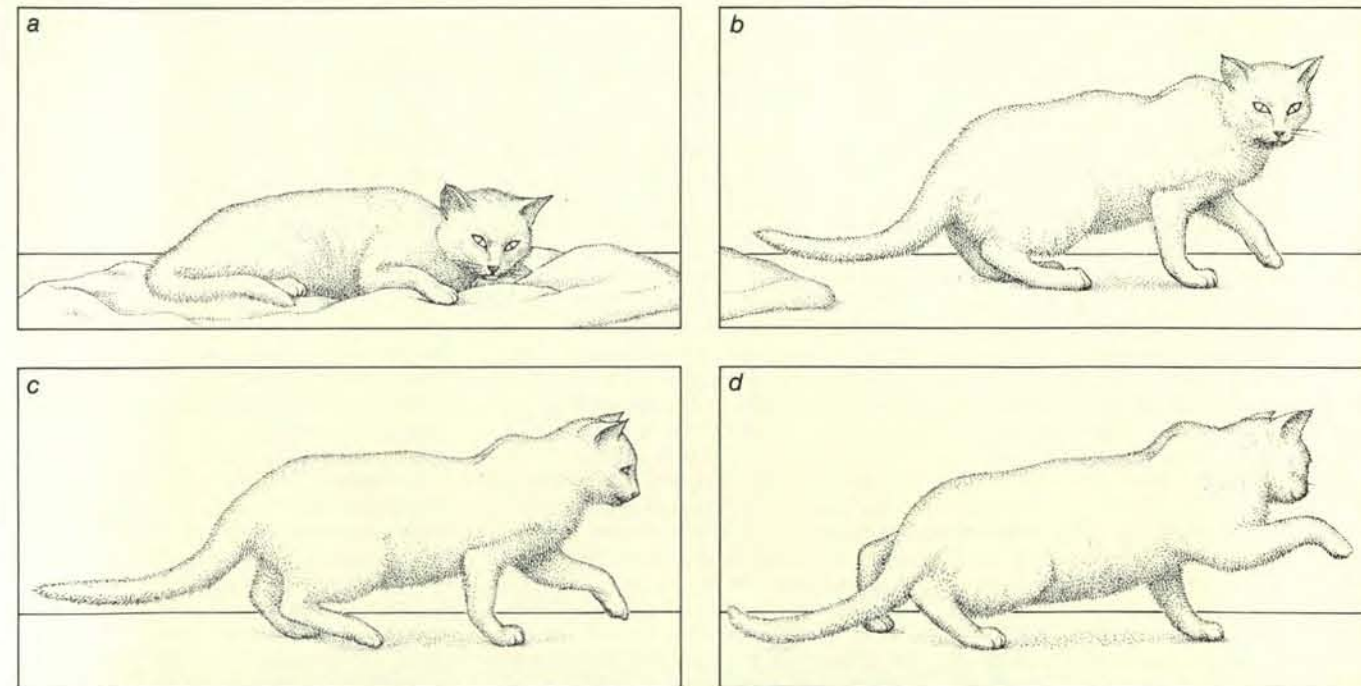
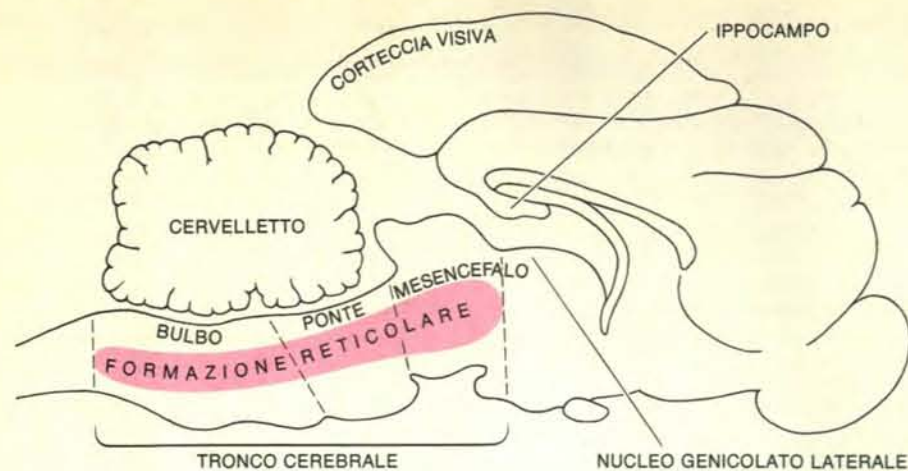
Oltre agli aspetti elettroencefalografici essenzialmente identici nel sonno REM e nella veglia, in questi due stati anche il tasso di attività dei neuroni è del tutto simile nella maggior parte delle subunità cerebrali. Una delle più sorprendenti somiglianze è stata identificata nell'ippocampo, che è filogeneticamente la parte più antica della corteccia cerebrale. Durante gran parte del sonno REM e della veglia, si registra nell'ippocampo un tracciato regolare con una frequenza di circa sette onde al secondo. Questo aspetto, detto «ritmo theta», è completamente differente dall'attività che l'ippocampo ha nel sonno a onde lente, meno regolare e con evidenti onde a punta. Un'ulteriore somiglianza si nota nella temperatura del cervello: dopo essere diminuita di una frazione di grado nel sonno a onde lente, aumenta circa al livello della veglia durante un episodio di sonno REM.

Quindi molte caratteristiche del sonno REM sono parallele a quelle della veglia e completamente diverse da quelle del sonno a onde lente. È da notare che tutti gli aspetti citati sopra sono segni di attivazione della formazione reticolare, o sistema reticolare attivatore, situato in profondità nel tronco cerebrale. Nel 1949, Giuseppe Moruzzi e Horace W. Magoun, lavorando insieme presso la Northwestern University, dimostrarono che tale formazione è responsabile del risveglio cerebrale: così, nel sonno REM, il cervello è decisamente sveglio. A causa della giustapposizione riposo-risveglio, il sonno REM è chiamato a volte «sonno paradosso».

La stretta connessione tra sonno REM e stato di veglia e di allerta è stata messa in rilievo di recente da una scoperta inaspettata, verificatasi nel mio laboratorio. Un mio allievo, Robert Bowker, stava lavorando sulle onde elettriche che appaiono «spontaneamente» (cioè senza stimolazione esterna evidente) durante il sonno REM. Queste onde brevi, di ampiezza elevata, sono dette punte PGO in relazione alle strutture cerebrali dove sono state maggiormente studiate: il ponte (dove si riteneva che esse originassero) e due parti del sistema ottico, il corpo genicolato laterale e la corteccia visiva (regione occipitale della corteccia cerebrale).

Si riteneva che le punte PGO fossero limitate quasi esclusivamente al sonno REM e che comparissero solo di rado nel sonno a onde lente. Un giorno, nel mio laboratorio, Bowker urtò incidentalmente la gabbia in cui era in corso di registrazione l'attività del cervello di un gatto in fase di sonno a onde lente. Quasi immediatamente comparve nel tracciato una punta PGO. Un'ulteriore ricerca mostrò che punte PGO potevano essere evocate sia nel sonno REM sia nel sonno a onde lente tramite suoni e stimoli tattili. Quindi le punte PGO, anziché eventi elettrici spontanei presenti solo nel sonno REM, apparvero come risposte di allerta multisensoriale, evocabili in più stati cerebrali.

La conclusione che tali punte sono risposte di allerta indusse Bowker a riesaminare le onde chiamate «potenziali di movimento oculare» (EMP dall'inglese *Eye Movement Potentials*) che compaiono nello stato di veglia. Nella registrazione elettrica tali onde sono identiche come aspetto alle punte PGO e si osservano nelle stesse strutture cerebrali. Si riteneva, invece, che la presenza di potenziali di movimento oculare dipendesse dal livello di illuminazione dell'ambiente in cui si trovava il gatto: tali potenziali non erano stati osservati, infatti, in gatti svegli al buio. La loro dipendenza dall'illuminazione era considerata una significativa differenza rispetto alle punte PGO.

Ritenendo che un gatto, in una camera di registrazione buia, potesse essere «annoiato», Bowker ne risvegliò l'interesse diffondendo attorno alla gabbia odore di tonno. Presto la registrazione rilevò potenziali di movimento oculare identici alle punte PGO. Emissioni brevi e improvvise di un suono intenso condussero allo stesso risultato. Ciò induce a

Il sonno REM senza paralisi di un gatto è illustrato in questi disegni ricavati da fotogrammi di un filmato, eseguito nel laboratorio dell'autore presso l'Università della Pennsylvania. La paralisi che normalmente accompagna il sonno REM è stata eliminata tramite una piccola lesione nel tronco cerebrale; il grado di paralisi varia con la lesione. Nell'episodio filmato, durante il sonno REM, il gatto solleva la testa e si raddrizza (*a*). Riesce a stare in piedi (*b*) e comincia a camminare (*c*). Poi, in un angolo del laboratorio, dà colpi nell'aria con le zampe (*d*). Nelle varie fasi, esso mostra tutti i segni di un normale sonno REM, compreso l'estendersi della membrana nittitante (la «terza palpebra») fino a coprire gran parte dell'occhio. L'aspetto del dorso leggermente insellato suggerisce che il tono non sia stato disinibito in tutti i muscoli.

Le strutture encefaliche interessate nel sonno sono evidenziate in questo schema di una sezione di encefalo di gatto. La parte frontale è a sinistra. Le lesioni che eliminano la paralisi nel sonno REM sono praticate nel ponte, una struttura del tronco cerebrale situata tra il bulbo e il mesencefalo.

pensare che sia le punte PGO sia i potenziali di movimento oculare siano forme particolari di una risposta di allerta generale. Tale risposta può essere evocata tramite stimolazione durante lo stato di veglia, durante il sonno a onde lente o durante il sonno REM. L'attivazione degli stessi neuroni può chiaramente avvenire anche in modo spontaneo durante il sonno REM. Così, durante un episodio di sonno REM, il cervello si risveglia raggiungendo uno stato che assomiglia allo stato di veglia e di allerta, anche se il contatto sensoriale con l'ambiente è molto ridotto.

Che cosa, durante il sonno REM, impedisce all'animale, il cui sistema nervoso centrale è sveglio e in allerta, ma per la maggior parte privo di contatto sensoriale con il mondo esterno, di nuocere a se stesso? Si può rispondere che durante ogni periodo REM l'azione dei neuroni motori del midollo spinale, che provocano la contrazione dei muscoli scheletrici (volontari), è inibita. Ne deriva che i muscoli sono atonici, cioè senza tono, e l'animale è paralizzato. (Va sottolineato che atonia e paralisi non sono sinonimi: una paralisi può derivare da varie condizioni oltre che dall'atonia; ma, se i muscoli sono atonici, l'animale risulta inevitabilmente paralizzato.)

Il grado di attività dei muscoli scheletrici è spesso registrato tramite un elettrodo inserito nel collo. I muscoli del collo appartengono al gruppo dei muscoli antigravitari, responsabili del mantenimento della stazione eretta. Come abbiamo visto, quando il gatto entra nella fase di sonno leggero, il suo tono muscolare diminuisce in quanto si riduce il numero di fibre attive in ciascun muscolo; quando, invece, esso entra nella fase di sonno REM, il tracciato relativo al tono muscolare diventa piatto, indicando una perdita completa di tono. Vi sono occasionali esplosioni di attività muscolare, però, che provocano contrazioni muscolari involontarie in varie parti del corpo. Nel mio laboratorio abbiamo lavorato con gatti nei quali l'atonia del sonno REM era stata eliminata tramite lesioni nel ponte praticate distruggendo una piccola massa di tessuto nervoso con un filo metallico rovente, introdotto secondo coordinate predeterminate.

Dopo aver praticato una lesione di questo tipo, si può osservare una notevole attività, che succede a un periodo di sonno a onde lente, subentrando nel momento in cui il gatto dovrebbe normalmente entrare nel sonno REM con atonia. Il gatto solleva la testa, raddrizza il corpo ed esegue movimenti alternati con gli arti. Tenta di stare in piedi e in alcuni casi ci riesce; in altri casi è addirittura in grado di camminare. Gatti con determinate lesioni mostrano altri comportamenti, che di solito si notano nella veglia. Eseguono movimenti tipici di orientamento verso la preda, di ricerca e di attacco, movimenti che sono raramente rivolti a oggetti presenti nell'ambiente.

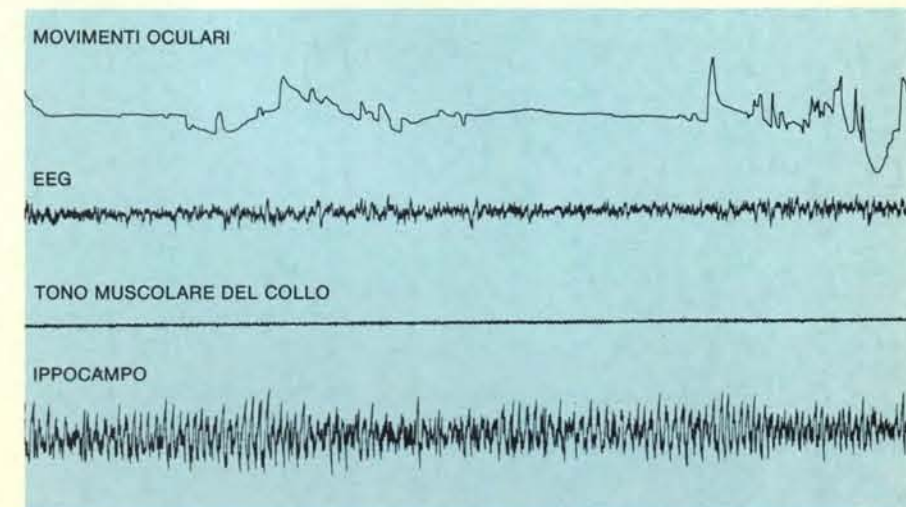
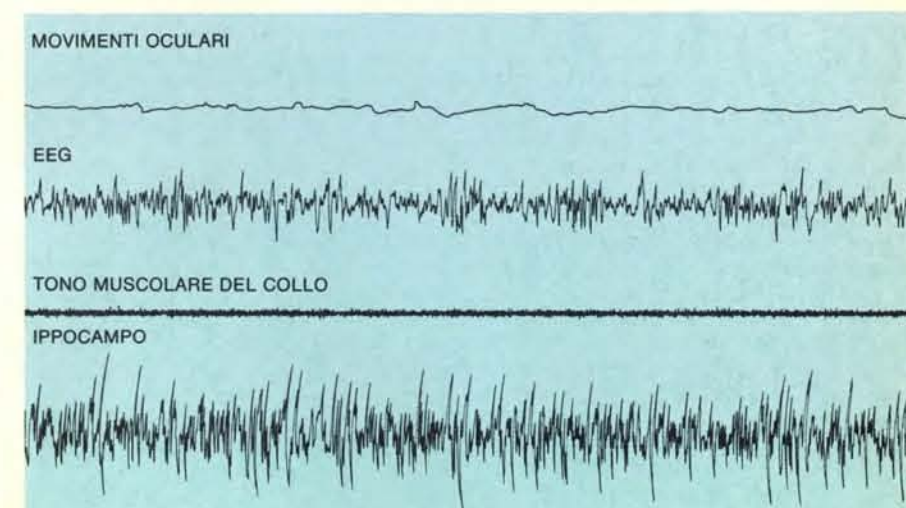
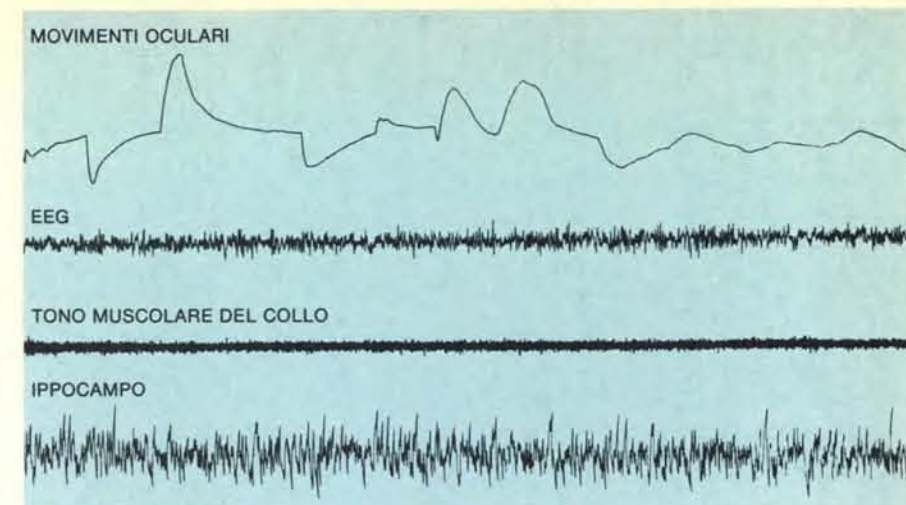
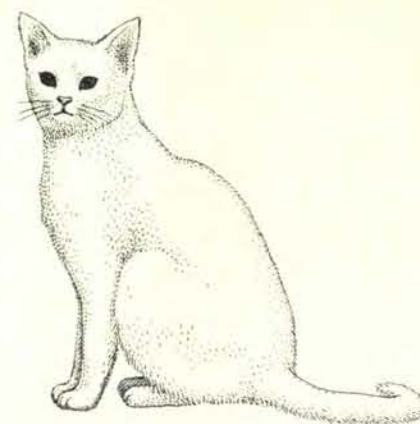
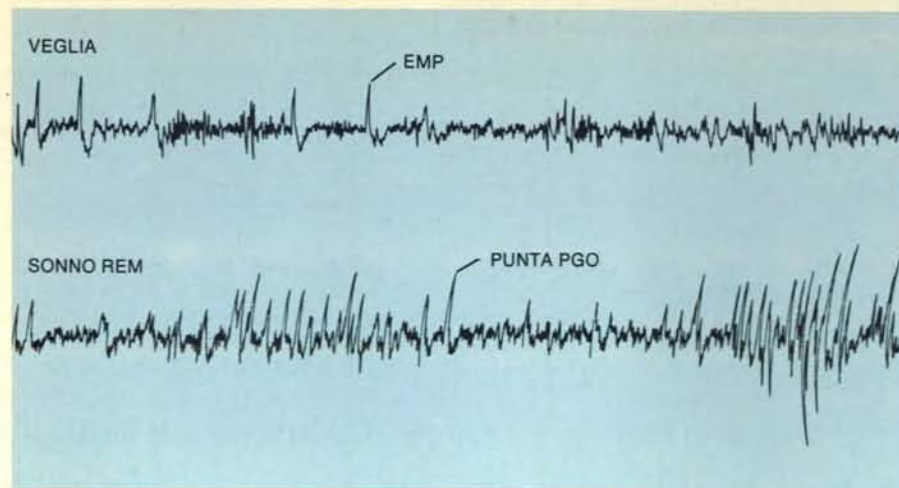
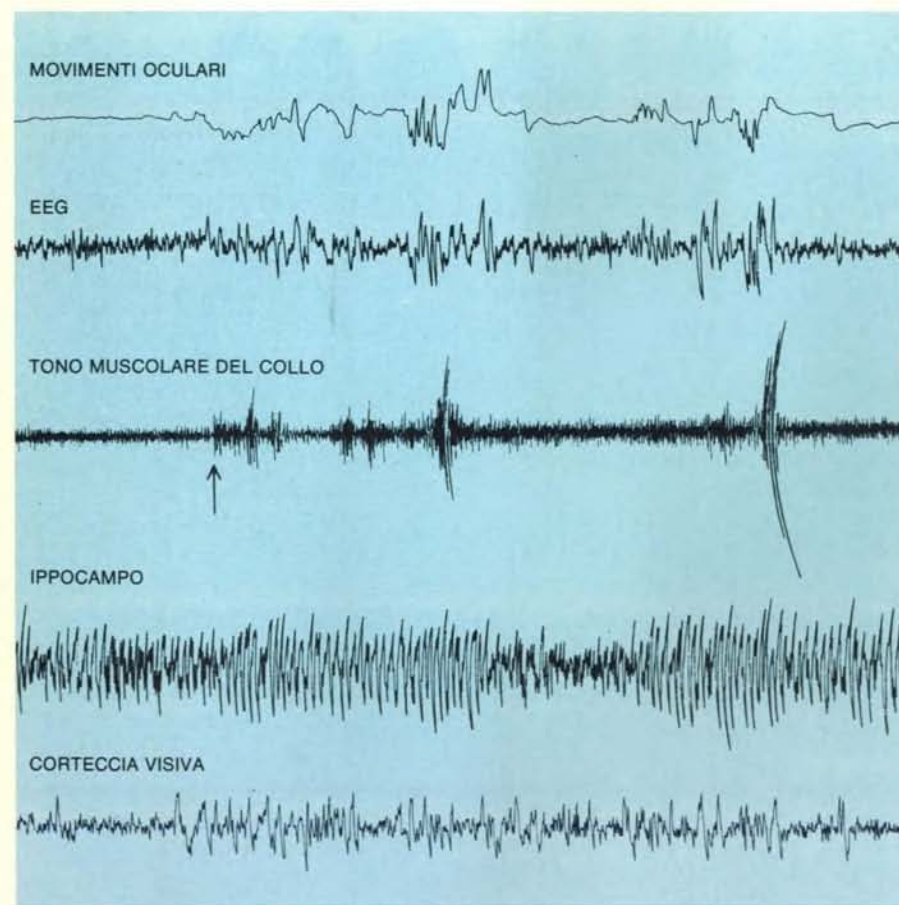
Come ho osservato, il tipo di attività muscolare che il gatto esibisce in questi straordinari episodi dipende dalla dimensione e dalla posizione della lesione cerebrale. Osservando al microscopio ottico sezioni di cervello di un gatto che poteva reggersi soltanto sulle zampe anteriori, sono state rilevate piccole lesioni, situate simmetricamente nella parte dorsale, o superiore, del ponte. In un gatto che poteva reggersi su tutte e quattro le zampe e che era in grado di camminare sono state osservate, invece, lesioni più estese in una posizione più ventrale, o più bassa del ponte. Un gatto che mostrava un comportamento aggressivo, colpendo con le zampe a più riprese il pavimento dinnanzi a sé, presentava lesioni che si estendevano in avanti fino al mesencefalo. Il significato di queste posizioni sarà ulteriormente chiarito quando verranno esaminate le vie nervose interessate nell'inibizione del movimento durante il sonno REM.

Parecchie fonti di informazione suggeriscono che gli insoliti movimenti osservati fanno parte effettivamente del sonno REM senza atonia. Il fenomeno è stato riconosciuto la prima volta da Michel Jouvet e François Delorme dell'Università di Lione. Come nel normale sonno REM, l'elettroencefalogramma dei gatti che si muovono durante il sonno è simile a quello registrato durante la veglia. La membrana nittitante copre parzialmente l'occhio e le pupille diventano sottili fessure. Le registrazioni a livello dell'ippocampo mostrano il ritmo theta. La mia allieva Joan Hendricks ha inoltre costretto i gatti a dormire in un'imbragatura imbottita che li immobilizzava; quelli con lesioni, che avrebbero dovuto impegnarsi in movimenti complessi se fossero stati liberi, mostrarono quei rapidi movimenti degli occhi, dei baffi e delle dita, che normalmente si osservano nel sonno REM. Durante questi episodi, come durante il normale sonno REM, la temperatura del cervello aumentava.

La Hendricks ha anche dimostrato che un aspetto particolarmente interessante del sonno REM è presente nel sonno REM senza atonia. Pier Luigi Parmiggiani e i suoi colleghi dell'Università di Bologna hanno dimostrato che nel gatto la capacità di regolare la temperatura corporea manca durante il sonno REM, nel quale questo animale diventa simile agli animali a sangue freddo (pesci, anfibi e rettili) e la Hendricks ha messo in rilievo che, durante il sonno REM senza atonia, i gatti non tremano e non arruffano il pelo in risposta al freddo, anche se mostrano tale risposta quando sono svegli. Pertanto, le lesioni a livello del ponte non ripristinano la capacità di rispondere al freddo.

Il mio lavoro attuale è prevalentemente orientato alla comprensione di come le piccole lesioni a livello del ponte determinino la perdita dell'atonia, propria del sonno REM, e la produzione di movimenti simili a quelli di un animale sveglio. Abbiamo concentrato la nostra attenzione sul ponte perché nel 1962, all'inizio dei suoi studi sul sonno, Jouvet aveva dimostrato che gatti, privati di tutta la parte del cervello situata anteriormente al ponte, mostravano ancora periodi di atonia e movimenti oculari rapidi identici a quelli del sonno REM. Nel corso dei miei studi ho concluso che vi sono almeno due sistemi neuronali, che operano nel sonno REM senza atonia: uno è responsabile della disinibizione del tono muscolare, l'altro della produzione di una «spinta motoria», un impulso generalizzato verso la locomozione.

Il primo sistema sembra il più semplice dei due. Sappiamo che, durante il sonno REM, è attivo un centro inibitorio nel bulbo rachidiano o midollo allungato. (L'azione dei neuroni può essere sia eccitatoria sia inibitoria. Le aree inibitorie riducono l'eccitabilità dei neuroni bersaglio.) Tale centro, dunque, durante il sonno REM, interferisce con i neuroni spinali che attivano i muscoli scheletrici, un meccanismo ritenuto responsabile della perdita di tono dei muscoli scheletrici.

Se il centro inibitorio del bulbo rachidiano fosse, a sua volta, sotto il controllo eccitatorio di una rete di cellule nervose

Nel gatto, tramite modificazioni della postura, del tono muscolare e dell'attività cerebrale, si possono distinguere tre stati: veglia, sonno a onde lente e sonno REM. Nella veglia *(in alto)* gli occhi e la testa dell'animale si muovono in risposta a stimoli acustici e visivi. L'elettroencefalogramma (EEG), che registra le modificazioni di tensione elettrica tra elettrodi posti sul cranio, presenta onde di bassa ampiezza e di alta frequenza. Nei muscoli scheletrici (volontari), c'è un notevole tono che viene misurato, in genere, nel collo. Quando il gatto rivolge la propria attenzione a un oggetto, si registra nell'ippocampo un tracciato regolare, detto ritmo theta. Quando esso entra nel sonno a onde lente *(al centro)*, si raggomitola e i movimenti oculari cessano; l'EEG mostra onde a bassa frequenza e ad ampiezza elevata. Il tono muscolare diminuisce e l'ippocampo presenta un ritmo irregolare. Quando, infine, il gatto entra in un periodo di sonno REM *(in basso)*, che si verifica all'incirca ogni 25 minuti durante il sonno, la sua posizione raggomitolata si rilassa leggermente e ricompaiono i movimenti oculari rapidi, un tracciato elettroencefalografico con onde a frequenza elevata e il ritmo theta. Il tono dei muscoli scheletrici scompare invece completamente.

Il risveglio del cervello durante il sonno REM è stato dimostrato di recente nel laboratorio dell'autore. Onde chiamate potenziali di movimento oculare (EMP) compaiono, nella veglia, nell'area visiva della corteccia cerebrale *(in alto)*. Un tempo si riteneva che tali potenziali dipendessero solo dal livello di illuminazione, ma Robert Bowker, allievo dell'autore, ha dimostrato che essi possono essere evocati da vari stimoli, persino al buio completo: sono risposte di allerta multisensoriali e sono onde identiche alle onde note come punte PGO, che compaiono spontaneamente nelle stesse aree del cervello, durante il sonno REM *(in basso)*. Nella fase REM il cervello è allerta, ma tale stato deriva dall'attività cerebrale più che da informazioni sensoriali.

Nei gatti con lesioni, episodi di movimento durante il sonno si verificano quando tutti i parametri, eccetto il tono muscolare, sono simili a quelli del sonno REM. Queste somiglianze hanno indotto l'autore a concludere che tali episodi indicano l'esistenza di un sonno REM senza atonia. Le registrazioni effettuate durante il sonno REM normale riguardano i movimenti oculari rapidi, il tracciato elettroencefalografico con onde di bassa ampiezza, il ritmo theta e le punte PGO. Questi segni sono tutti presenti nelle registrazioni eseguite, come quella in figura, durante gli episodi di movimento che conseguono a un danno nel ponte. La differenza notevole è che nel sonno REM senza atonia i muscoli scheletrici del gatto hanno riguadagnato gran parte del tono, cosicché l'animale può muoversi. La freccia indica l'inizio di un episodio di movimento durante il sonno REM.

situata nel ponte, un danno a livello di quest'ultimo potrebbe interrompere la via eccitatoria. L'azione del suddetto centro verrebbe interrotta e i muscoli scheletrici manterrebbero il loro tono anche durante il sonno REM. L'estensione della disinibizione del tono, e quindi il numero di muscoli che i gatti potrebbero usare, dipenderebbe dalla sede delle lesioni.

Questa ipotesi ha ricevuto conferma dai risultati ottenuti da Jouvet e dai suoi colleghi di Lione, i quali hanno dimostrato che un fascio di fibre nervose si origina proprio dorsalmente all'area del ponte dove abbiamo prodotto le lesioni. Tale fascio termina poi direttamente a livello dell'area inibitoria del bulbo. Abbiamo provocato lesioni sia all'origine sia lungo il suo decorso e ciò ha prodotto gatti che, durante il sonno REM, avevano un certo tono nei muscoli del collo, pur potendo sollevare solo di poco la testa. Il che suggerisce che altri neuroni, probabilmente nella formazione reticolare, devono essere lesionati per disinibire completamente il tono muscolare.

Alcune prove sperimentali sostengono, quindi, l'ipotesi che, alla base dell'atonia del sonno REM, vi sia una connessione abbastanza diretta tra ponte, centro inibitorio del bulbo e muscoli scheletrici. Sapendo che il sonno REM nei soggetti umani è associato a un'intensa attività onirica, si sarebbe propensi a concludere che, interrompendo questa connessione, sarebbe possibile vedere l'animale che mima i suoi sogni. Oltre alle difficoltà che emergono nell'attribuire stati mentali complessi ad altre specie, vi sono anche buone ragioni per pensare che questa non sia una spiegazione completa di ciò che accade negli episodi di sonno REM in assenza di atonia.

Le funzioni del cervello si basano su strutture anatomiche strettamente interconnesse tra loro. Sarebbe irragionevole aspettarsi che l'unico risultato che deriva dalla lesione di un'area centrale del tronco cerebrale sia la distruzione di strutture che inibiscono i neuroni motori. In effetti, abbiamo la prova che sistemi diversi da quello inibitorio presente nel bulbo sono danneggiati dalle lesioni provocate a livello del ponte.

La prova più importante riguarda il livello totale di attività locomotoria presente quando i gatti lesionati sono svegli. Essi, nel sonno a onde lente o nella veglia, non mostrano un anomalo incremento del tono muscolare: l'effetto sul tono muscolare si manifesta soltanto nel sonno REM. Sembrano, però, in generale più attivi di quanto non lo fossero prima di subire le lesioni. Quando vengono liberati nel laboratorio, possono anche arrecarsi molestia, correndo qua e là come se inseguissero qualcosa in modo incoerente.

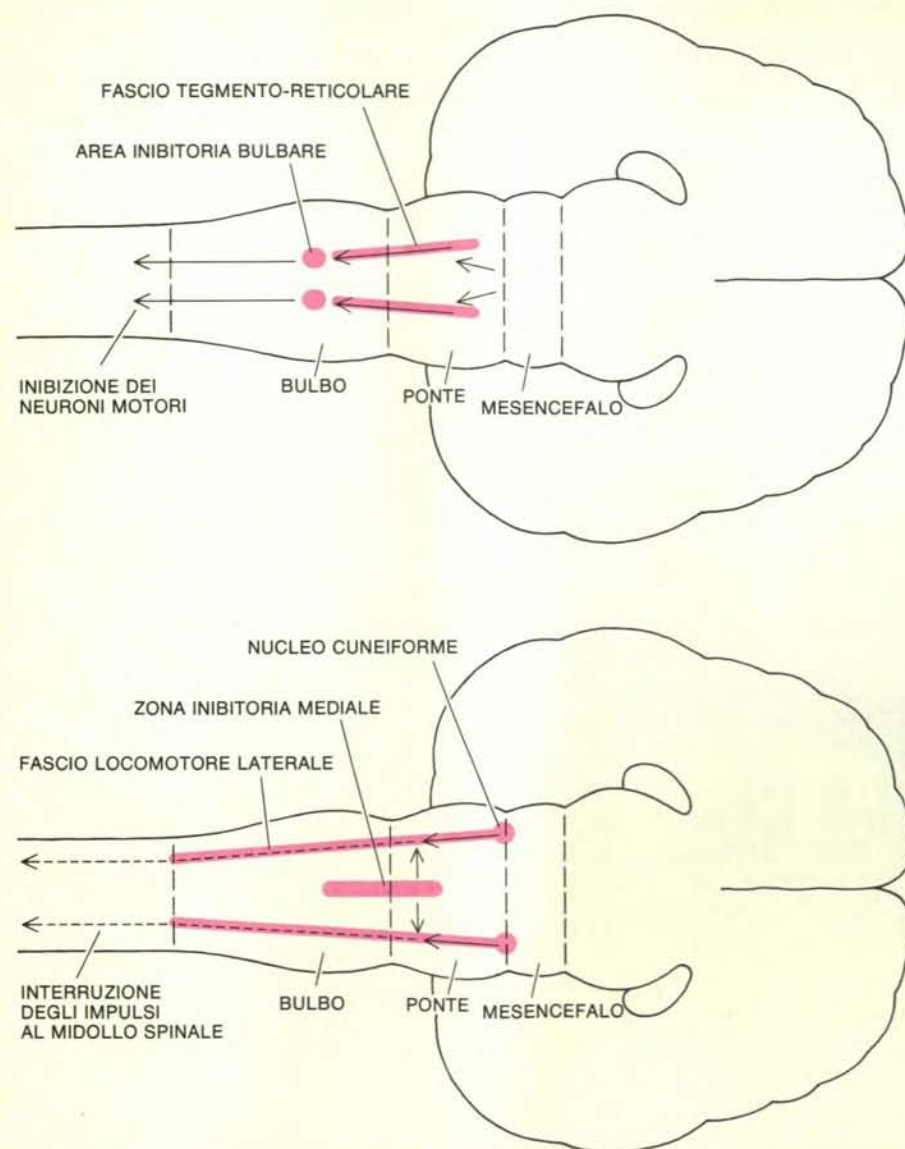
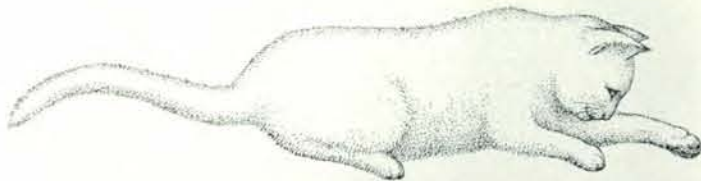
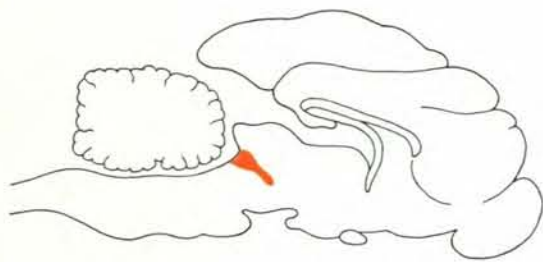
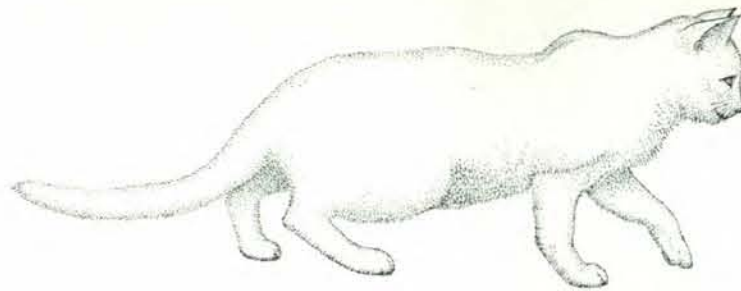
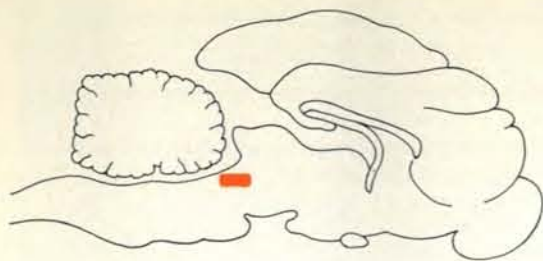
Per confermare l'impressione di un'attività locomotoria insolita, abbiamo eseguito delle prove in libertà. Parecchi soggetti sono stati esaminati prima e dopo aver subito le lesioni e sono stati liberati in una stanza il cui pavimento era suddiviso in quadrati. Il numero di quadrati sui

quali ogni gatto passava in un periodo di 30 minuti è stato utilizzato come misura della sua attività locomotoria. Tutti i gatti che presentavano sonno REM senza atonia hanno mostrato anche incrementi di attività dopo essere stati lesionati: tali incrementi variavano dal 30 al 261 per cento ed erano tutti statisticamente significativi. Questa scoperta mi ha portato a concludere che le lesioni a livello del ponte interessano la fonte da cui ha origine la spinta motoria generalizzata, fonte che è probabilmente un sistema anatomico diverso da quello che regola il tono muscolare, poiché quando gli animali erano svegli il loro tono muscolare non era influenzato dalle lesioni.

Il nostro lavoro sta cominciando a fornire un quadro di come potrebbe agire il secondo sistema di inibizione durante il sonno. La mia ipotesi su questo funzionamento deriva in prevalenza dalla ricerca eseguita sul controllo nervoso della locomozione. È stato dimostrato che la regione del tronco cerebrale a cui siamo interessati può regolare la locomozione senza nessun contributo da parte dei centri cerebrali superiori: questa regolazione coinvolge tre strutture. Un «generatore della locomozione» nel midollo spinale comprende neuroni che controllano e coordinano i movimenti reciproci degli arti nella deambulazione e nella corsa. Esso è poi sotto il controllo eccitatorio di una seconda struttura: un «centro della locomozione» nel tronco cerebrale. Se questo centro viene stimolato, stimola a sua volta il generatore dei movimenti reciproci. La terza struttura è una regione di controllo nel ponte, connessa al centro della locomozione tramite una via inibitoria. Quando questa regione viene attivata, sopprime l'attività del centro della locomozione. Così, nel sonno REM, il ponte potrebbe indirettamente sopprimere, assieme al tono muscolare, la spinta motoria. D'altro canto, un danno al ponte potrebbe disinibire i muscoli utilizzati nella locomozione.

L'esistenza di un centro della locomozione nel tronco cerebrale e di un generatore della locomozione nel midollo spinale è stata postulata da molti ricercatori, ma la loro organizzazione anatomica non è ancora nota. La prima prova dell'esistenza di un centro della locomozione è stata ottenuta da M. L. Shik, F. V. Severin e G. N. Orlovskii dell'Università statale di Mosca; i loro risultati sono stati poi confermati da altri ricercatori. Shik e i suoi colleghi hanno dimostrato che l'applicazione di una corrente elettrica a un sistema neuronale che ha origine nel nucleo cuneiforme, situato nella regione caudale del mesencefalo, può indurre dei gatti privati degli emisferi cerebrali a muoversi su una passatoia mobile e persino a trottare o a galoppare. Io sospetto che sia questo il centro che le lesioni a livello del ponte disinibiscono durante il sonno REM.

Un recente studio di Shigemi Mori e dei suoi collaboratori presso il Collegio di medicina di Ashikawa in Giappone sostiene l'esistenza del centro della locomozione e suggerisce in che modo esso può essere connesso al ponte. I ricercatori giapponesi hanno scoperto che i fasci capaci di indurre la locomozione in gatti privati degli emisferi cerebrali discendono dal mesencefalo fino alla parte più bassa del tronco cerebrale, sotto forma di due sottili cordoni che decorrono lateralmente nel ponte e nel bulbo. Essi hanno anche dimostrato che vi sono, nella parte centrale del tronco cerebrale, aree in grado di facilitare o di inibire l'azione del centro della locomozione.

Il centro della locomozione viene eccitato o inibito a seconda dell'area esatta che è stimolata. La stimolazione elettrica della zona facilitatoria più ventrale è in grado di indurre movimenti reciproci nelle zampe posteriori dell'animale senza nessuna diretta stimolazione del centro della locomozione. La stimolazione dorsale riduce, invece, il tono muscolare e inibisce la locomozione. Le zone facilitatoria e inibitoria scoperte da Mori e dai suoi colleghi sono situate presso la linea mediana del tronco cerebrale e sono vicine alle regioni del ponte dove sono state

Reti neuronali distinte sembrano responsabili dell'inibizione del tono muscolare e della spinta motoria nel sonno REM. L'azione dei neuroni sulle cellule bersaglio può essere sia inibitoria sia eccitatoria. Nel bulbo, un'area riconosciuta come inibitoria sui neuroni motori svolge una parte importante nel ridurre il tono muscolare *(in alto)*. Il fascio tegmento-reticolare connette il ponte con il centro inibitorio. Nel sonno REM il ponte è attivato e a sua volta eccita il centro bulbare tramite questa e altre vie. Il centro bulbare inibisce i neuroni motori e fa aumentare l'atonia. Un fascio locomotore laterale ha un ruolo importante nel ridurre la spinta motoria *(in basso)*. Esso fuoriesce in basso dal tronco cerebrale e si connette con le strutture del midollo spinale, potendo indurre movimenti negli arti dei gatti in cui gli emisferi cerebrali, anteriormente al mesencefalo, sono stati rimossi. In profondità, nel tronco cerebrale, vi sono neuroni che possono eccitare o inibire il fascio locomotore: essi sono organizzati in una zona facilitatoria e in una zona inibitoria, situate nello stesso piano verticale presso la linea mediana dell'encefalo. Nel sonno REM il ponte stimola la zona inibitoria, rendendo inattivo il fascio locomotore e annullando la spinta motoria. Così, un danno a livello del ponte può eliminare, in entrambi i sistemi, il controllo inibitorio.

La posizione delle lesioni nel cervello del gatto influisce notevolmente sul tipo di movimento che l'animale mostra durante il sonno REM senza atonia. Un gatto con una piccola lesione in alto nel ponte può reggersi solo sulle zampe anteriori *(in alto)*. Il danno può interrompere solo parzialmente l'effetto della zona inibitoria mediale sul fascio locomotore mostrato nella figura precedente. Un gatto con una lesione più ampia e più profonda nel ponte si regge su tutte e quattro le zampe e cammina *(al centro)*; il danno cioè può eliminare completamente l'effetto della zona inibitoria. Un gatto con una lesione che si estende nel mesencefalo mostra un comportamento aggressivo *(in basso)*: il danno in questo caso interferisce con le vie nervose che regolano il comportamento aggressivo e che provengono dagli emisferi cerebrali.

praticate le nostre lesioni. In questo modo la connessione tra il ponte e il presunto centro della locomozione sta diventando più chiara.

Questi risultati servono a facilitare l'interpretazione degli effetti delle lesioni nel ponte. Per esempio, nel gatto citato prima, che poteva stare eretto sulle zampe e camminare, la lesione aveva eliminato pienamente l'azione della zona inibitoria mediale sul centro della locomozione. Nel gatto, invece, che poteva sollevarsi solo sulle zampe anteriori, la lesione non aveva interrotto completamente l'azione della zona inibitoria mediale. La mia collaboratrice Graziella Mann e io abbiamo scoperto di recente che un ulteriore danno alla zona inibitoria può far sì che un gatto, capace di sollevarsi solo sulle zampe anteriori, riesca a camminare anche su tutte e quattro le zampe durante il sonno REM senza atonia. Tuttavia, le connessioni anatomiche all'interno del ponte e tra il centro della locomozione nel tronco cerebrale e il generatore della locomozione nel midollo spinale rimangono ipotetiche e sono necessari ulteriori studi in questo campo. Nei gatti che mostrano un comportamento aggressivo vi sono altre complicazioni: questi animali presentano sempre danni a livello delle vie nervose che discendono dagli emisferi cerebrali e che sono interessate nel controllo dell'aggressione durante la veglia.

Non si deve pensare, tuttavia, che quando le connessioni anatomiche specifiche saranno note, il problema del controllo dei muscoli nel sonno REM potrà considerarsi completamente risolto. Potrebbero, infatti, esservi coinvolti meccanismi anche più sottili di quelli illustrati sopra. Per esempio, René Drucker-Colin dell'Università del Messico con i suoi allievi Gloria Arankowsky, Raul Aguilar e io, abbiamo recentemente dimostrato che le scariche neuronali ad alta frequenza, che si verificano in molte regioni del cervello durante il sonno REM, sono necessarie per disinibire l'attività nel sonno REM senza atonia.

La somministrazione di un comune antibiotico, il cloramfenicolo, in una dose simile a quella utilizzata per curare infezioni batteriche, provoca la ricomparsa dell'atonia in gatti con lesioni nel ponte, che altrimenti avrebbero mostrato episodi di movimento durante il sonno REM. Quei gatti sembrano normali durante la veglia. Drucker-Colin aveva dimostrato in precedenza che il cloramfenicolo riduce la frequenza di scarica nei neuroni della formazione reticolare. Non è noto il meccanismo tramite il quale l'attività neuronale viene ridotta. Il cloramfenicolo è noto però come inibitore della sintesi proteica e questa azione potrebbe essere responsabile della riduzione della frequenza di scarica dei neuroni. Qualunque sia, comunque, il meccanismo d'azione del farmaco, dopo la sua somministrazione i gatti con le lesioni possiedono troppo poca spinta locomotoria per produrre quell'attività complessa che mostrerebbero altrimenti durante il sonno REM.

La ricerca eseguita finora sul sonno REM senza atonia indica che c'è un legame tra l'elevata reazione di risveglio durante il sonno REM e la riduzione dell'attività motoria coordinata. L'organismo è ovviamente ben protetto da un meccanismo che gli impedisce di muoversi quando il cervello è molto attivo, ma non risponde alle stimolazioni esterne. Non è necessario invocare argomentazioni teleologiche per spiegare l'esistenza di un nesso tra risveglio del cervello e inibizione motoria: esso potrebbe esistere persino nello stato di veglia.

Una persona o un animale, posti di fronte a uno stimolo nuovo o inatteso, normalmente esitano un poco prima di agire. Quasi tutti noi abbiamo sperimentato un momento di esitazione o persino una sensazione di cedimento delle ginocchia quando abbiamo visto un'automobile avanzare a grande velocità verso di noi. Nella veglia tale inibizione motoria è in generale di breve durata; tuttavia fa pensare che persino in questo stato vi sia un nesso tra aumento momentaneo della vigilanza e riduzione dell'attività motoria.

Lo studio di questo legame potrebbe avere almeno una conseguenza clinica importante. Il disturbo chiamato narcolessia è caratterizzato da improvvise e imprevedibili cadute dallo stato di veglia direttamente nel sonno REM o dallo stato di veglia in una paralisi senza perdita di coscienza. È interessante notare che un forte stimolo (come quello che accompagna la collera, il riso, la sorpresa o il rapporto sessuale) è la causa più frequente delle crisi di narcolessia. Può darsi che negli individui che ne sono affetti vi sia una soglia anormalmente bassa per il legame tra reazione di risveglio e inibizione motoria; gli stimoli che comunemente causano una reazione di risveglio potrebbero così portare, in momenti inadeguati, ad atonia o al sonno REM completo.

Oltre alle informazioni sulle possibili conseguenze cliniche, lo studio del sonno REM senza atonia ne fornirà probabilmente altre sulla natura dello stesso sonno REM. Esso è già stato utile nel chiarire un problema fondamentale. Come abbiamo visto, nella maggior parte delle regioni cerebrali, i neuroni seguono lo stesso comportamento nella transizione dallo stato di veglia a quello di sonno. Dalla veglia al sonno a onde lente si verifica una riduzione del tasso di attività e quindi, quando comincia un episodio di sonno REM, si nota un ritorno al tasso dello stato di veglia. Tra i neuroni che mostrano un differente comportamento vi sono quelli dell'area del ponte, nota come nucleo del rafe dorsale: queste cellule utilizzano la serotonina come trasmettitore per alterare l'attività dei neuroni a cui sono connesse. La velocità con cui avviene la scarica neuronale diminuisce da circa due impulsi al secondo nello stato di veglia a quasi zero nel sonno REM, passando attraverso un livello intermedio nel sonno a onde lente.

Era stata avanzata l'ipotesi che l'inattività dei neuroni nel nucleo del rafe dorsale fosse una caratteristica fondamentale del sonno REM. Infatti si era supposto che la diminuzione dell'attività neuronale provocasse l'inizio di tale fase. Per provarlo ho collaborato con Barry L. Jacobs e Michael Trulson della Princeton University, i quali avevano già studiato a fondo l'attività neuronale del rafe dorsale sia nello stato di veglia sia nel sonno normale. Quando è stata registrata l'attività di queste cellule nel sonno REM senza atonia, i risultati sono stati sorprendenti. Nei gatti con lesioni nel ponte, i neuroni del rafe, invece di diventare inattivi, hanno aumentato nuovamente la loro attività dopo il sonno a onde lente. Tale incremento ha portato a una velocità di circa un impulso al secondo e quindi non a quella dei neuroni nello stato di veglia. Ha condotto, comunque, a una velocità che è circa sei volte maggiore di quella che si nota nel normale sonno REM.

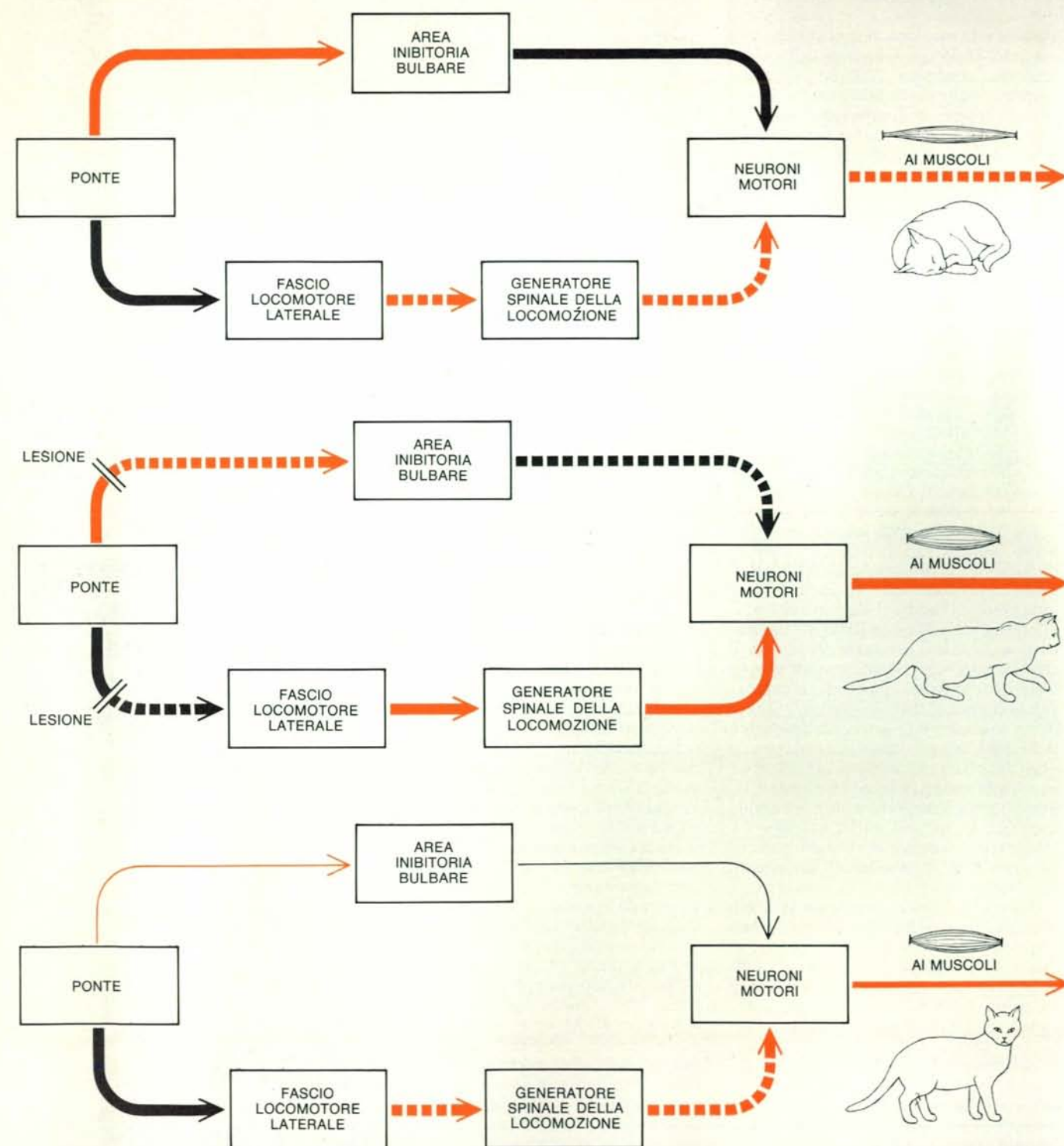
Vi sono almeno due spiegazioni plausibili per questo inatteso risultato. La prima è che l'attività muscolare insolita che si osserva nel sonno REM senza atonia in qualche modo faccia tornare l'informazione al ponte, così da eccitare i neuroni del nucleo del rafe dorsale. La spiegazione in alternativa è che le lesioni a carico del ponte interessano un meccanismo motorio più centrale che normalmente, nel sonno REM, inibisce i neuroni del rafe. La prima ipotesi deve dipendere dall'arrivo di impulsi dalla periferia al centro del sistema nervoso, mentre la seconda chiama in causa solo eventi che si verificano nel cervello.

Per mettere alla prova le due ipotesi Jacobs e allievi hanno utilizzato due farmaci: la succinilcolina e il carbacolo. La succinilcolina è un farmaco chimicamente molto simile al curaro. Agisce alla giunzione tra un neurone motore e la cellula muscolare che questo neurone attiva e determina una paralisi temporanea interrompendo la connessione a questo livello. Quando è stata iniettata in gatti normali svegli, i neuroni del rafe hanno mantenuto l'attività che avevano prima dell'iniezione. Quindi questi neuroni potevano essere attivi sia nella veglia con paralisi sia nel sonno REM senza atonia e quindi senza paralisi. Abbiamo così concluso che, essendo attivi negli animali svegli e paralizzati, essi non venivano stimolati, nel sonno REM senza atonia, da informazioni provenienti dalla periferia quale risultato di un'attività muscolare vigorosa.

Il carbacolo provoca paralisi agendo direttamente sul meccanismo sopra descritto, responsabile dell'atonia. Dopo l'iniezione, ha reso quasi completamente inattivi i neuroni del rafe. L'inattività dei neuroni del rafe dorsale sembra così un fenomeno collaterale, più che una parte fondamentale del sonno REM o addirittura un elemento del meccanismo che dà origine a un episodio di sonno REM. Essa sembra correlata con l'inibizione motoria centrale del sonno REM, che può essere revocata tramite una lesione a livello del ponte. Di recente il mio allievo Peter Reiner ha studiato l'altro gruppo di cellule nervose che diventano inattive durante il sonno REM. Questi neuroni sono situati presso l'area dove abbiamo praticato le lesioni e utilizzano la noradrenalina come neurotrasmettitore. I suoi risultati preliminari suggeriscono che anche l'inattività di queste cellule è correlata con l'inibizione motoria del sonno REM.

Tramite tecniche sperimentali, che ci permettono di separare la paralisi muscolare associata al sonno profondo dalla attività del cervello, stiamo apprendendo molte cose su ciò che è fondamentale nel sonno REM e su ciò che, invece, rappresenta solo un evento collaterale. Le ipotesi correnti sulla paralisi che si instaura durante il sonno REM richiederanno senza dubbio ulteriori studi per essere pienamente accettate: in particolare, l'ipotesi avanzata sopra sulle vie separate per l'inibizione del tono muscolare e per l'inibizione della spinta motoria e l'ipotesi che questa seconda via sia attiva, in qualche modo, anche durante la veglia.

Sia nel caso che queste idee vengano dimostrate valide, sia nel caso che vengano smentite, il sonno REM senza atonia con-

tinuerà a essere una tecnica sperimentale preziosa. Oltre a rendere possibile la separazione di ciò che, nel sonno REM, è essenziale da ciò che è collaterale, esso potrebbe costituire il mezzo con cui studiare un comportamento complesso come quello dell'aggressione, che è evocato in genere come risposta a stimoli esterni, ma potrebbe essere generato nel sonno REM, indipendentemente dall'ambiente che circonda l'animale. Inoltre, grazie alle somiglianze tra veglia e sonno REM, il confronto tra questi due stati continuerà senza dubbio a essere interessante. Nei movimenti dell'animale in stato di veglia riconosciamo una correlazione con i suoi stati interni. Fino a ora questa correlazione ci mancava per il sonno REM. L'abbiamo acquisita per il sonno REM senza atonia ed essa sembra, in un certo senso, una finestra aperta sul cervello durante il sonno.

I sistemi che inibiscono il movimento durante il sonno REM possono operare anche in stati di emergenza fisica durante la veglia. Essi sono rappresentati in forma schematica nelle tre parti della figura. Le connessioni eccitatorie sono indicate in colore, quelle inibitorie in nero. Durante il normale sonno REM, il ponte attiva spiccatamente il centro inibitorio del bulbo *(in alto)*. La sua zona inibitoria mediale agisce, invece, sul fascio locomotore laterale: ne risulta una completa paralisi. Nel sonno REM senza paralisi, le lesioni a livello del ponte tagliano le connessioni tra esso e il fascio locomotore e l'area bulbare *(al centro)*. Il gatto può muoversi, ma la disinibizione del tono muscolare è generalmente incompleta. Nella veglia, quando viene percepito uno stimolo inaspettato e minaccioso, a volte si verifica un'inibizione transitoria del movimento *(in basso)*: probabilmente il risultato di una riduzione della spinta motoria, essendovi scarsa perdita di tono muscolare.

TEMI METAMAGICI

di Douglas R. Hofstadter

Un discorso sul Lisp si conclude con un pantagruelico banchetto all'italiana

Il mese scorso ho descritto il rompicapo della Torre di Brahma, ideato da Edouard Lucas: si tratta di trasferire 64 dischi d'oro da un perno di diamante a un altro, servendosi di un terzo perno sul quale si possono momentaneamente sistemare i dischi. I dischi devono venir sfilati e mossi uno per volta e non è permesso sistemare un disco sopra uno più piccolo. Il problema che vi avevo posto era quello di arrivare a una definizione ricorsiva, espressa come funzione in Lisp, della strada seguita per raggiungere l'obiettivo (e, quindi, secondo i sacerdoti di Brahma, far finire il mondo).

Ho precisato che la ricorsività è abbastanza evidente: per trasferire 64 dischi da un perno all'altro, servendosi di un terzo, è sufficiente sapere come trasferire 63 dischi da un perno all'altro, servendosi di un terzo. Ricapitolando, supponete di avere, all'inizio, la Torre di Brahma coi 64 dischi sul perno *a*. L'illustrazione di questa pagina rappresenta schematicamente i 64 dischi con quattro soltanto. Prima di tutto trasferite, basandovi sulla vostra presunta abilità di spostare 63 dischi, 63 dischi dal perno *a* al perno *c*, servendovi del perno *b* come «perno di supporto». Si può vedere il risultato dell'operazione nella seconda figura a partire dall'alto. (Si noti che, nella prima illustrazione, 4 sta per 64, 3 sta per 63, ma per qualche motivo 1 non sta per 61. Non è curioso?) Ora semplicemente sfilate l'unico disco rimasto su *a*, il più grande di tutti, e infilatelo sul piano *b*, come si vede nella terza figura a partire dall'alto.

Vi accorgerete ora di come sia semplice arrivare alla fine: usate semplicemente di nuovo la vostra capacità di trasferire la fila di dischi da *c* di nuovo su *b*, questa volta usando *a* come perno di supporto. Si noti come in questa manovra il perno *a* svolga il ruolo ausiliario precedentemente ricoperto da *c* per lo spostamento dei 63 dischi. Un attimo prima che l'ultimo disco venga messo a posto, la situazione è quella che si può vedere nella figura in basso. Perché prima e non dopo che l'ultimo disco venga messo a posto? Perché allora l'intero mondo si polverizzerebbe ed è troppo difficile disegnare la polvere!

Il rompicapo della Torre di Brahma con quattro dischi

Qualcuno a questo punto potrebbe obiettare che ho trascurato la parte più difficile: «Hai semplicemente postulato una magica capacità di spostare 63 dischi!» Anche se così può sembrare, non c'è niente di magico nel postulato. Dopo tutto per spostare 63 dischi, dovete semplicemente sapere come spostarne 62 e, per spostarne 62, dovete semplicemente saperne spostare 61. Così procedendo si discende la fila fino ad arrivare al «caso embrionale» del rompicapo della Torre di Brahma, il rompicapo con un solo disco. Capisco che seguire il processo passo passo può essere un po' noioso, ma è semplice routine. In linea di principio ora potreste effettivamente espletare l'intero processo, se foste disposti a vedere la fine del mondo!

Come prima approssimazione a una funzione del Lisp scriviamo una descrizione del metodo in linguaggio comune. Chiamiamo i tre perni «pp», «pa» e «ps», cioè, rispettivamente, perno di partenza, perno d'arrivo e perno di supporto. Avremo:

```
Per spostare una torre di altezza n
 da pp a pa, servendosi di ps:
  se n = 1,
     allora spostare semplicemente l'unico disco
     direttamente da pp a pa;
  altrimenti
     compiere i tre passi seguenti:
       (1) spostare una torre di altezza
       n – 1 da pp a ps
           servendosi di pa;
       (2) portare un disco da pp a pa
       (3) spostare una torre di altezza
       n – 1 da ps a pa,
           servendosi di pp.
```

I passi 1 e 3 sono le due chiamate ricorsive; scivolando quasi nel paradosso, sembrano implicare realmente l'abilità che stanno cercando di definire. La caratteristica che le salva dal paradosso è che parlano di n – 1 dischi invece che di n dischi. Si noti che al passo 1 ps ricopre il ruolo di «arrivo», mentre pp svolge il ruolo di «supporto», mentre al passo 3 ps svolge il ruolo di «partenza», mentre pp svolge il ruolo di «supporto». Dato che l'intero processo è ricorsivo, i perni si scambieranno molte volte i ruoli durante i trasferimenti. Questa è la bellezza del rompicapo e, in un certo senso, è anche la bellezza della ricorsività.

Come si passa dal linguaggio comune al Lisp? È abbastanza semplice:

```
(def spostare-la-torre
 (lambda (n pp pa ps)
  (cond
   ((eq n 1) (portare-un-disco pp pa))
   (t(spostare-la-torre(sub 1 n) pp ps pa)
     (portare-un-disco pp pa)
     (spostare-la-torre(sub 1 n) ps pa
pp)))))
```

Dove sono gli equivalenti in Lisp delle parole «da», «a» e «servendosi di»? Sembrano scomparse! Allora come fa il genietto del Lisp a sapere quale ruolo ricopre a ciascuno stadio ciascun perno? La risposta è che l'informazione viene fornita *dalla posizione*. In questa definizione di funzione vi sono quattro parametri: un numero intero e tre «perni fittizi», il primo dei quali è la partenza, il secondo l'arrivo e il terzo è il supporto. Nell'elenco iniziale dei parametri (dopo il «lambda») si trovano nell'ordine «pp, pa, ps». Nella prima chiamata ricorsiva, la traduzione in Lisp del passo 1, si trovano nell'ordine «pp, ps, pa», da cui risulta che ps e pa si sono scambiati i ruoli. Nella seconda chiamata ricorsiva, la traduzione in Lisp del passo 3, si può vedere che ps e pp si sono scambiati i ruoli.

Il fatto è che i nomi di atomo «pp», «pa» e «ps» non hanno alcun significato intrinseco per il genietto. Avrebbero potuto essere «ananas», «banana» e «ciliegia» che sarebbe stato lo stesso. I loro significati vengono definiti operativamente a seconda del posto in cui compaiono nelle varie parti della definizione della funzione. Sarebbe stato un grosso errore quindi scrivere, per esempio, «(spostare-la-torre(sub1 n) pp pa ps)» come traduzione in Lisp del passo 1, perché non dà nessuna indicazione del fatto che ps c pa, in quel passo, devono scambiarsi i ruoli.

Rimane senza risposta una domanda importante. Che cosa succede quando il simpatico genietto del Lisp arriva al passo che chiede «spostare-un-disco»? Si precipita immediatamente al tempio di Benares e solleva realmente un pesante disco d'oro? Oppure, prosaicamente, solleva un disco di plastica che sta su un tavolo e lo trasferisce da un perno di plastica a un altro sempre di plastica? In altre parole, si verifica un'azione reale invece di una pura computazione?

```
(ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca bc
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ca bc ba ca bc ab ac bc ba ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab ac bc ba ca bc ab ac bc ab ca cb ab ca bc ba ca cb
 ab ac bc ab ca cb ab)
```

Il genietto del Lisp fornisce le mosse per risolvere un rompicapo della Torre di Brahma di altezza 9

In teoria è possibilissimo. Anche in pratica l'esecuzione di un comando in Lisp potrebbe realmente portare un braccio meccanico a spostarsi fino a un luogo dato, prendere una qualsiasi cosa la sua mano meccanica possa afferrare, portare quell'oggetto in un altro luogo dato e poi lasciarlo andare. In tempi di robot industriali non si tratta certo di fantascienza. Se non vi sono, però, braccia e mani meccaniche, che significato può avere la cosa?

Una possibilità ovvia e apparentata alla precedente è che su uno schermo compaia una visualizzazione del rompicapo e che «portare-un-disco» faccia comparire un'*immagine* di una mano che si muove, *sembra* afferrare un disco, prenderlo e sistemarlo da qualche altra parte. In tal modo si simula graficamente il rompicapo, il che può essere fatto con vari gradi di realismo.

Supponiamo, tuttavia, di non avere a disposizione un hardware o un software grafici raffinati, supponiamo di volere solamente creare una formula scritta che ci dice come muovere le nostre morbide e organiche mani umane per risolvere il rompicapo. Allora, nel caso del rompicapo con 3 dischi, la formula desiderata potrebbe essere «ab ac bc ab ca cb ab» o indifferentemente «12 13 23 12 31 32 12». Che cosa potrebbe aiutarci a raggiungere questo più umile scopo?

Se avessimo un programma per muovere un braccio, non ci interesserebbe il valore risultante, ma la serie strutturata di «effetti collaterali» che produrrebbe. Qui, al contrario, ci interessa solo il valore che ci darà il programma: una lista strutturata di atomi. La lista per n = 3 è stata costruita a partire da due liste per n = 2. L'idea era stata presentata alla fine dell'articolo del mese scorso:

ab ac bc   ab   ca cb ab

In Lisp, per dividere i gruppi, invece di usare spazi più grandi, usiamo le parentesi. Quindi il nostro scopo, per n = 3, potrebbe essere quello di produrre la lista sandwich «((ab ac bc) ab (ca cb ab))». Un sistema per formare una lista a partire dai suoi componenti è quello di usare ripetutamente «cons». Quindi, se i valori degli atomi «ananas», «banana», «ciliegia» sono 1, 2, 3, allora il valore risultante da «(cons ananas (cons banana (cons ciliegia nil)))» sarà la lista «(1 2 3)». Esiste, tuttavia, un sistema più rapido per ottenere lo stesso risultato e precisamente scrivere «(list ananas banana ciliegia)». Si ottiene lo stesso valore. Analogamente, se gli atomi «pp» e «pa» sono legati, rispettivamente, ad «a» e «b», allora l'esecuzione del comando «(list pp pa)» darà «(a b)». «List» è una strana funzione perché si applica a un numero qualsiasi di argomenti (anche a nessun argomento: in tal caso «(list)» dà come valore nil!).

Affrontiamo ora il problema del valore che vogliamo ci dia la funzione «spostare-un-disco». Ha due parametri che rappresentano perni e vorremmo che ci portasse a un singolo atomo ricavato dai nomi di quei perni, come «ab» o «12». Per il momento sarà più semplice postulare che i nomi dei perni siano i numeri 1, 2 e 3. In tal caso, per ottenere il numero 12 da 1 e 2, basterà fare un po' di aritmetica: moltiplicare il primo numero per 10 e aggiungere il secondo. Ecco il Lisp che ci viene incontro:

```
(def spostare-un-disco
 (lambda (pp pa)
  (plus (times 10 pp) pa)))
```

Se invece i nomi dei perni sono atomi non numerici possiamo usare una funzione standard del Lisp chiamata «concat» che si applica ai valori dei propri argomenti (un numero qualsiasi, come per «list») e li concatena tutti in modo da formare un grosso atomo. Allora «(concat 'con'ca-t'e'nate)» dà l'atomo «concatenate». In tal caso possiamo scrivere

```
(def spostare-un-disco
 (lambda(pp pa) (concat pp pa)))
```

Abbiamo comunque risolto metà della riga finale del problema della torre.

L'altra metà del problema sta nel valore che darà la parte ricorsiva. La cosa è semplicissima: vogliamo semplicemente che ci dia una lista a sandwich in cui i valori delle due chiamate ricorsive si affianchino al valore della singola chiamata di spostare-un-disco. Possiamo così modificare di poco la nostra precedente definizione ricorsiva aggiungendo la parola «list»

```
(dcf spostare-la-torre
 (lambda (n pp pa ps)
  (cond
   ((eq n 1) (spostare-un-disco pp pa))
   (t
    (list
     (spostarc-la-torre (sub 1 n) pp ps pa)
     (spostare-un-disco pp pa)
     (spostare-la-torrc (sub 1 n) ps pa
pp))))))
```

Contrattiamo ora un piccolo scambio con il genietto del Lisp

```
—> (spostare-la-torre 4 'a 'b 'c)
(((ac ab cb) ac (ba bc ac))
          ab
((cb ca ba) cb (ac ab cb)))
—>
```

Straordinario! Funziona davvero! Non è bella? Il mese scorso ho chiamato questa lista «brahma».

Supponiamo di voler eliminare tutte le parentesi interne cosicché rimanga solo una lunga successione ininterrotta di atomi. Potremmo, per esempio, ottenere, invece dell'intricato brahma, «(ac ab cb ac ba bc ac ab cb ca ba cb ac ab cb)». Fornirebbe minore informazione, ma sarebbe più efficace nella sua opacità. In questo caso non ci serviamo della funzione «list» per formare il nostro sandwich di tre valori, ma usiamo un'altra funzione che elimini le parentesi dai due valori ricorsivi laterali.

```
n=0   oeso
n=1   oeso
        (e solo oeso)
n=2   oeso
        (e solo oeso)
         (e solo oeso (e solo oeso))
n=3   oeso
        (e solo oeso)
         (e solo oeso (e solo oeso))
          (e solo oeso (e solo oeso) (e solo oeso (e solo oeso)))
```

Una breve tabella delle espansioni ricorsive della variabile del Lisp «oeso»

In questo caso giunge a proposito la funzione standard del Lisp «append». Essa unisce qualsiasi numero di liste, facendo cadere le loro parentesi più esterne. Allora «(append '(a (b)) '(c) nil '(de))» porta alla lista di cinque elementi «(a (b) c d e)» invece che alla lista di quattro elementi «((a (b)) (c) nil (d e))» cui si sarebbe arrivati se si fosse applicata la funzione «list» invece di «append». Usando «append» e una versione un po' modificata di «spostare-un-disco» a cui applicarla, possiamo formulare una definizione finale di «spostare-la-torre» che fa al caso nostro:

```
(def spostare-la-torre
 (lambda (n pp pa ps)
  (cond
   ((eq n 1) (spostare-un-disco pp pa))
   (t
    (append
     (spostare-la-torre (sub 1 n) pp ps pa)
     (spostare-un-disco pp pa)
     (spostare-la-torre (sub 1 n) ps pa
pp))))))

(def spostare-un-disco
 (lambda (pp pa) (list (concat pp pa))))
```

Per verificare la cosa ho chiesto al genietto del Lisp di risolvere il rompicapo della Torre di Brahma per una torre alta 9 dischi. Nella pagina precedente si può vedere quello che, quasi istantaneamente, mi ha risposto.

Ci siamo finora occupati di un esempio di ricorsività raffinato e difficile. Passiamo ora ad un tipo di ricorsività che ci mette alla prova in modo diverso. Questo tipo di ricorsività si riallaccia a un'osservazione marginale che avevo fatto nell'articolo precedente. Ho usato lo strano nome di variabile «oeso», precisando che si tratta di un acronimo ricorsivo che sta per «oeso (e solo oeso)». Per questo motivo si può sviluppare «oeso» quante volte si vuole. L'unico principio da rispettare è che ogni occorrenza di «oeso» a un livello dato deve essere rimpiazzata dalla frase composta di due parti «oeso (e solo oeso)» per passare al livello successivo. Si può vedere come la cosa avvenga nell'illustrazione in alto in questa pagina. La prova per noi consiste nello scrivere una funzione in Lisp che dia il valore di «oeso» dopo n sviluppi ricorsivi, per un n qualsiasi. È irrilevante che, per n maggiore di 3, lo sviluppo diventi enormemente lungo: siamo dei teorici!

C'è solo un problema: qualsiasi funzione del Lisp deve dare come valore un'unica struttura del Lisp (un atomo o una lista), mentre gli ingressi della nostra tabella non soddisfano tale criterio. Per esempio, l'ingresso per n = 2 è costituito da un atomo seguito da due liste. Per eliminare questo inconveniente possiamo trasformare ciascun ingresso in una lista racchiudendola entro un paio di parentesi più esterne. Ora il nostro scopo è coerente con il Lisp. Come ci siamo arrivati?

Il ragionamento ricorsivo ci dice che la riga terminale, ossia il caso embrionale, occorre quando n = 0 e che altrimenti l'ennesima riga è data dalla riga precedente in cui sia stato rimpiazzato l'atomo «oeso», dovunque esso occorra, con la lista «(oeso (e solo oeso))», senza le parentesi più esterne. Possiamo scriverlo:

```
(def sviluppo-di-oeso
 (lambda (n)
  (cond
   ((eq n 0)' (oeso))
   (t
    (rimpiazza
     'oeso
     '(oeso (e solo oeso))
     (sviluppo-di-oeso (sub 1 n)))))))
```

Non abbiamo però specificato che cosa intendiamo per «rimpiazza». Dobbiamo essere molto cauti nel definire come eseguire la nostra operazione di rimpiazzamento. Prendete una qualsiasi riga della tabella di «oeso» e vedrete che contiene un elemento in più rispetto alla riga precedente. Perché? Perché l'atomo «oeso» viene rimpiazzato ogni volta con una lista a due elementi, le cui parentesi, come ho puntualizzato prima, sono cadute nel momento del rimpiazzamento. Il punto arduo è proprio questo della scomparsa delle parentesi. Un esempio di rimpiazzamento con caduta delle parentesi meno complicato di quello ricorsivo di «oeso» potrebbe essere «(rimpiazza 'a (1 2 3) '(a b a))» che avrebbe come valore «(1 2 3 b 1 2 3)». Questo tipo di rimpiazzamento comporta, piuttosto che l'esatta sostituzione di un atomo con una lista, l'unione interna di una lunga lista.

Cerchiamo di specificare, in Lisp, attraverso la ricorsività, come al solito, che cosa intendiamo per «rimpiazzamento» di tutte le occorrenze dell'atomo «atm» con una lista «lyst», all'interno di una lunga lista «lungalista». È un bel rompicapo! Permettetemi di darvi un suggerimento: tenete conto del fatto che la risposta per «(a b a)» è costruita sulla risposta per «(ba)». Osservate anche altri casi semplici analoghi, risalendo fino al caso embrionale.

Si dà il caso embrionale quando «lungalista» è nil. Allora non succede niente e la nostra risposta dovrà essere nil.

Il caso ricorsivo comporta che si costruisca una risposta più complessa a partire da una più semplice che si assume come data. Torniamo al nostro esempio «(a b a)». Abbiamo costruito la risposta complessa «(1 2 3 b 1 2 3)» a partire dalla risposta più semplice, «(b 1 2 3)», aggiungendo a essa, mediante la funzione append, «(1 2 3)». Anche «(b 1 2 3)», d'altra parte, può essere considerata una risposta complessa risultante dalla risposta più semplice «(1 2 3)», inserendo «b» mediante la funzione cons. Perché nel primo caso ci serviamo di append e nel secondo di cons? Semplicemente perché nel primo caso si tratta dell'atomo «a» che *deve* essere sostituito, mentre nel secondo si tratta dell'atomo «b» che *non deve* essere sostituito. Quest'ultima operazione ci permette di tentare una definizione ricorsiva di «rimpiazza» nel modo seguente:

```
(def rimpiazza
 (lambda (atm lyst lungalista)
  (cond
   ((null lungalista) nil)
   ((eq (car lungalista) atm)
    (append
     lyst
     (rimpiazza atm lyst (cdr lungali-
sta))))
   (t
    (cons
     (car lungalista)
     (rimpiazza atm lyst (cdr lungali-
sta)))))
))
```

Come potete vedere, c'è un caso embrionale (in cui lungalista è uguale a nil) e poi un caso ricorsivo in cui compare «append» e un caso ricorsivo in cui compare «cons». Verifichiamo ora questa definizione con un nuovo esempio.

```
—> (rimpiazza 'a '(1 2 3) '(a (a) b a))
(1 2 3 (a) b 1 2 3)
—>
```

Ops! È andato *quasi* bene: la funzione ha mancato completamente una delle occorrenze di «a». Vuol dire che nella nostra definizione di «rimpiazza» abbiamo dimenticato qualcosa. In effetti, se tornate indietro, vi accorgerete che ci siamo lasciati sfuggire un'asserzione indimostrata, precisamente quella che gli elementi di lungalista siano sempre atomi. Abbiamo trascurato la possibilità che lungalista possa contenere sottoliste. Che fare in tal caso? Operare il rimpiazzamento anche all'interno delle sottoliste delle sottoliste e così via. Riuscite a immaginare un modo per giungere alla definizione?

Abbiamo già osservato in precedenza un processo ricorsivo in cui tutte le parti a tutti i livelli di una struttura dovevano essere esplorate: si trattava della funzione «atomcount» in cui abbiamo applicato un processo ricorsivo contemporaneamente sia a «car» sia a «cdr». La riga ricorsiva era «(plus (atomcount (car s)) (atomcount (cdr s)))». Anche qui si tratterà di una cosa del genere. Avremo una riga ricorsiva caratterizzata da istruzioni del tipo «rimpiazza», una da applicarsi al car di lungalista e l'altra da applicarsi al cdr di lungalista, invece di una sola da applicarsi al cdr. La cosa, se ci pensate, è perfettamente sensata. Supponete di voler sostituire tutti gli unicorni d'Europa con porscopini. Un modo per perseguire questo nefando obiettivo sarebbe quello di dividere in due l'Europa: il Portogallo (il car dell'Europa) e il resto dell'Europa (il cdr dell'Europa). Dopo aver rimpiazzato tutti gli unicorni del Portogallo con porscopini e anche tutti gli unicorni del resto dell'Europa con porscopini, potete finalmente rimettere insieme i due pezzi di un'Europa riunificata (il che comporta un'operazione «cons»). Ovviamente, per portare a termine questa ignobile operazione anche per il Portogallo deve aver luogo un'analoga separazione con successivo ricongiungimento, e così via. Si potrebbe quindi avere:

```
(cons (rimpiazza 'unicorno
                 '(porscopino)
                 (car unità-geografica))
      (rimpiazza 'unicorno
                 '(porscopino)
                 (cdrunità-geografica)))
```

Ovvero, più in generale:

```
(cons (rimpiazza atm lyst (car lungalista))
      (rimpiazza atm lyst (cdr lungalista)))
```

Questa riga «cons» si applicherà sia al caso in cui lungalista non è un atomo sia a quello in cui è un atomo, diverso, però, da atm. Perché la cosa funzioni dobbiamo ampliare leggermente il caso embrionale: diremo che quando lungalista non è una lista, ma un atomo, «rimpiazza» non ha nessun effetto su lungalista; in questo modo comprendiamo anche la prima riga «null», cosicché possiamo tralasciarla. Mettendo insieme il tutto, otteniamo una nuova, migliore definizione:

```
(def rimpiazza
 (lambda (atm lyst lungalista)
  (cond
   ((atom lungalista) lungalista)
   ((eq (car lungalista) atm)
    (append
     lyst
     (rimpiazza atm lyst (cdr lungalista))))
   (t
    (cons
     (rimpiazza atm lyst (car lungalista))
     (rimpiazza atm lyst (cdr lungali-
sta)))))
))
```

Quando, ora, diciamo al genietto del Lisp «sviluppo-di-oeso 2», ci darà la lista «(oeso (e solo oeso) (e solo oeso (e solo oeso)))».

Bene, bene. Non è un magnifico risultato? Se non ci sembra proprio magnifico, possiamo forse compiere un passo ulteriore. Potrebbe risultare divertente creare un acronimo ricorsivo, cioè un acronimo contenente una lettera che sta al posto dell'acronimo stesso, ma che dire degli acronimi *reciprocamente* ricorsivi? Questo potrebbe voler dire, per esempio, avere due acronimi ciascuno dei quali contiene una lettera che sta al posto dell'altro acronimo. Potremmo ad esempio avere la coppia di acronimi «NOODLES» e «LINGUINI» che stanno per:

«NOODLES (oodles of delicious LINGUINI),
elegantly served»

[La frase inglese ha più o meno il seguente significato: «taglierini, una gran quantità di deliziosi linguini, elegantemente serviti». Per questo acronimo e i successivi, che riportiamo nella versione inglese dell'autore, si darà la traduzione italiana tra parentesi quadre. *n.d.t.*]

«luscious itty-bitty NOODLES got
usually in Naples, Italy»
[profumati, sottili taglierini che si fanno di solito a Napoli in Italia]

Si noti che «NOODLES» non è solo indirettamente ricorsivo, ma lo è anche direttamente e la cosa non crea problemi.

In generale, la nozione di ricorsività reciproca ha a che vedere con un sistema costituito da un numero qualsiasi di strutture intrecciate, ciascuna delle quali è definita nei termini di uno o più membri del sistema (fra i quali vi può essere anche la struttura stessa). Se parliamo di una famiglia di acronimi reciprocamente ricorsivi, intendiamo una collezione di parole in ciascuna delle quali le lettere stanno per una parola della famiglia.

Devo ammettere che questa idea di acronimi reciprocamente ricorsivi non ha alcuna particolare utilità pratica. Ciononostante, è abbastanza utile come divertente esempio di un fenomeno astratto abbastanza comune. Chi non si è almeno qualche volta divertito di fronte all'inevitabile circolarità delle definizioni del dizionario? Tutti si possono rendere conto che, alla fine, tutte le parole sono definite mediante un insieme fondamentale di termini che non è ulteriormente riducibile ma semplicemente si ripiega continuamente su se stesso. Potete divertirvi a cercare sul dizionario una parola qualunque e a rimpiazzare le parole della definizione di tale parola con le *loro* definizioni. [Nel testo che segue, le definizioni sono la traduzione letterale delle definizioni inglesi trovate da Hofstadter, non una ricostruzione nel dizionario italiano. *n.d.t.*] Io, una volta, ho applicato la cosa alla parola «amore» definita come «un forte effetto o attaccamento o devozione per una o più persone». Sostituendo «forte», «affetto», «attaccamento», «devozione» e «persona» sono arrivato a questa mescolanza: «una moralmente potente tendenza o stato mentale, che ha forza di carattere o di volontà, o un'affettuosa considerazione o lealtà, fedeltà o profondo affetto per uno o più esseri umani, in quanto distinti da cose o animali inferiori».

Non ancora soddisfatto, mi sono spinto oltre e ho ottenuto: «Un insieme di circostanze o attributi che caratterizzano una persona o una cosa in un dato momento, consciamente o inconsciamente, come un'unità piena di ossia avente una specifica abilità ossia capacità in un modo che ha a che vedere, tratta, ossia è capace di distinguere tra un buon e un cattivo comportamento ossia un'inclinazione ad agire o a muoversi in una particolare direzione, con l'attributo o la qualità di essere forte moralmente ossia considerazione, attenzione ossia interesse pieno di tenerezza ossia la qualità, lo stato di essere fedeli a quelle persone o ideali che uno si sente di difendere o sostenere ossia la condizione, qualità o stato di essere degno di fiducia o profondamente considerato ossia un tenero sentimento per una o più creature ossia caratteristica di una o più persone che vivono o esistono, in particolare in quanto separate o differenziate da ciò che si può pensare o a cui ci si può riferire come a una singola entità o da qualsiasi organismo vivente inferiore in rango, dignità o autorità, capace di muoversi ma non di prodursi da solo il cibo per fotosintesi». Non è romantico?

Torniamo agli acronimi reciprocamente ricorsivi. Mi sono sforzato di elaborarne una famiglia e, con mia sorpresa, ho scoperto che essi riguardavano per lo più, anche se non esclusivamente, il cibo italiano. Tutto cominciò con la scelta della parola «tomato» (pomodoro) e con la successiva decisione di usarla al plurale ottenendo

TOMATOES on MACARONI
(and TOMATOES only),
exquisitely SPICED
[pomodori sugli spaghetti (e pomodori soli), squisitamente insaporiti]

Le parole in lettere maiuscole inserite in parentesi sono quelle che sono anche acronimi. Ecco tutta la mia famiglia reciprocamente ricorsiva:

– MACARONI:
MACARONI and CHEESE

(a REPAST of Naples, Italy)
[spaghetti: spaghetti e formaggio, un pasto napoletano, in Italia]
– REPAST:
rather extraordinary PASTA
and SAUCE, typical
[pasto: una pasta al sugo del tutto straordinaria, piatto tipico]
– CHEESE:
cheddar, havarti, Emmentaler
(particularly SHARP Emmentaler)
[formaggio: formaggio duro dolce, Emmental (un Emmental particolarmente forte)]
– SHARP:
strong, hearty and rather pungent
[forte, robusto e piuttosto acuto]
– SPICED:
sweetly pickled in CHEESE
ENDIVE dressing
[insaporiti: leggermente salati, con formaggio, conditi in insalata]
– ENDIVE:
egg NOODLES, dipped
in vinegar eggnog
[taglierini all'uovo, immersi in eggnog all'aceto]
– NOODLES:
NOODLES (oodles of delicious
LINGUINI), elegantly served
[taglierini: gran quantità di deliziosi linguini, elegantemente serviti]
– LINGUINI:
LAMB CHOPS
(including NOODLES)
got usually in northern Italy
[linguini: cotolette di agnello (comprendenti taglierini) che si mangiano spesso nell'Italia del Nord]
– PASTA:
PASTA and SAUCE (that's ALL!)
[pasta: pasta al sugo (è tutto!)]
– ALL:
a luscious lunch
[tutto: un saporito pranzo]
– SAUCE:
SHAD and unusual COFFEE
(eccellente!)
[salsa: pesce e insolito caffé (eccellente!)]
– SHAD:
SPAGHETTI, heated al dente
[pesce: spaghetti cotti al dente]
– SPAGHETTI:
standard PASTA, always good, hot
particularly (twist, then ingest)
[spaghetti: pasta comune, sempre buona, particolarmente calda (arrotolare e mangiare)]
– COFFEE
choice of fine flavors,
particularly ESPRESSO
[caffé: scelta di buoni sapori, in particolare l'espresso]
– ESPRESSO:
excellent, strong, powerful, rich ESPRESSO,
suppressing sleep
outrageously
[espresso: un eccellente, forte, potente, ricco espresso che fa passare il sonno del tutto]
– BASTA!:
belly all stuffed (tummy ache!)
[pancia piena (mal di stomaco!)]
– LAMB CHOPS
LASAGNE and meatballs,
casually heaped onto
– PASTA SAUCE
[costolette di agnello: lasagne e polpette casualmente sistemate sul sugo della pasta]
– LASAGNE:
LINGUINI and SAUCE
and GARLIC
(NOODLES everywhere!)
[lasagne: linguini e salsa e aglio (taglierini ovunque!)]
– RHUBARB:
RAVIOLI, heated under butter
and RHUBARB (BASTA!)
[rabarbaro: ravioli scaldati al burro e rabarbaro (basta!)]
– RAVIOLI:
RIGATONI and vongole in oil,
lavishly introduced
[ravioli: rigatoni e vongole all'olio, aggiunto abbondantemente]
– RIGATONI:
rich Italian GNOCCHI
and TOMATOES
(or NOODLES instead)
[rigatoni: ricchi gnocchi e pomodori italiani (o invece taglierini)]
– GNOCCHI:
GARLIC NOODLES
ower crisp CHEESE,
heated immediately
[gnocchi: taglierini all'aglio, sopra formaggio grattugiato, scaldati immediatamente]
– GARLIC:
green and red LASAGNE
in CHEESE
[Lasagne verdi e rosse al formaggio]

Qualsiasi cuoco può vedere che non ci si è molto sforzati a far corrispondere a ciascun termine una frase che lo definisca correttamente: si tratta di una associazione più o meno arbitraria.

Che succede se miriamo a espandere alcune parole, per esempio, «pasta»? Dapprima abbiamo semplicemente «PASTA and SAUCE (that's ALL)». Al passo successivo «PASTA and SAUCE (that's ALL!) and SHAD and unusual COFFEE (eccellente!) (that's a luscious lunch)». Potremmo ovviamente continuare a sviluppare acronimi all'infinito, o almeno fino a riempire l'universo fino all'orlo con descrizioni fiume di piatti italiani. Ma, se fossimo meno ambiziosi e desiderassimo semplicemente riempire una mezza paginetta? In che modo potremmo arrestare questa ricorsività apparentemente senza fondo?

Le parole chiave sono «senza fondo», e la risposta che implicano è: mettete in atto un meccanismo che faccia arrivare a un fondo la ricorsività. La mancanza di un fondo deriva dal fatto che a ogni stadio ogni acronimo può svilupparsi, cioè può generare ulteriori acronimi. Se invece controllassimo rigidamente il procedimento di diffusione degli acronimi essendo generosi nei confronti delle prime poche «generazioni» e lasciando invece che un numero via via minore di acronimi diffonda progenie alle generazioni successive? Avverrebbe qualcosa di simile a una sequoia in una foresta che comincia con un solo «ramo» (il suo tronco) e quel ramo spande «progenie», cioè la prima generazione di rami più piccoli, che a loro volta spandono ulteriore progenie - ma alla fine il processo giunge a un fondo naturale per il fatto che i ramoscelli più piccoli non possono ulteriormente generare. Si possono fare analoghe considerazioni per gli «alberi» di acronimi reciprocamente ricorsivi. Se nello sviluppare «tomatoes» prendessimo sempre le stesse decisioni di controllo su quali acronimi sviluppare e quando, allora ci sarebbe uno e un solo tipo di sviluppo di «rhubarb». Se invece lasciamo spazio alla casualità, possiamo ottenere molte generazioni di «rhubarb», che presenterebbero fra loro una somiglianza precisa e rivelatrice, ma molto più sfuggente.

Che cosa possiamo fare? L'ideale è servirsi di un generatore di numeri casuali, l'equivalente a livello di calcolatore del lanciare in aria una moneta o tirare un dado. Faremo dipendere la decisione di sviluppare o meno un dato acronimo dal risultato di un ideale lancio di moneta. Ai primi stadi di sviluppo metteremo le cose in modo che risulti molto probabile che esca testa (sviluppo); agli stadi più avanzati croce (nessuno sviluppo). Per ottenere questo impiegherò la funzione del Lisp «rand». Non si applica a nessun argomento e ogni volta che è chiamata dà un numero reale compreso tra 0 e 1, in modo imprevedibile. (È un'esagerazione; in realtà è prevedibile al 100 per cento se si sa come calcolarlo, ma, dato che l'algoritmo è piuttosto oscuro, nella maggior parte dei casi la funzione si comporta in modo così irregolare da sembrare completamente casuale. Il tema della generazione di numeri casuali è molto affascinante e richiederebbe di per sé un articolo.)

Se vogliamo che un evento si verifichi con la probabilità del 60 per cento, dapprima chiediamo un valore a caso. Se il valore è 6 o inferiore a 6, andiamo avanti, altrimenti no. Dato che sul lungo periodo la funzione rand distribuisce uniformemente i suoi valori sull'intervallo tra 0 e 1, il procedimento ci darà il benestare a proseguire nel 60 per cento dei casi.

Questo per quanto riguarda le decisioni casuali. Come possiamo fare sì che un acronimo si sviluppi al nostro comando? Non è troppo difficile. Supponiamo che ogni acronimo sia una funzione del Lisp, come nell'esempio seguente:

```
(def tomatoes
 lambda ( )
  '(tomatoes on macaroni
    (and tomatoes only)
    exquisitely spiced))))
```

La funzione «tomatoes» non si applica a nessun argomento e semplicemente dà nuovamente la lista di parole in cui si sviluppa. Nulla di più semplice.

Supponiamo ora di avere una variabile «acronimo» il cui valore sia un particolare acronimo, ma non si sa quale. Come possiamo ottenere che quell'acronimo si sviluppi? Il modo in cui abbiamo composto l'acronimo deve fungere da chiamata di funzione. Perché un atomo possa chiamare una funzione, deve essere il car di una lista come negli esempi «(plus 2 2)», «(rand)» e «(rhubarb)». Se scrivessimo «(acronimo)», il genietto del Lisp prenderebbe l'atomo «acronimo» letteralmente per un nome di funzione - e questo sarebbe un fraintendimento. Non è certamente l'atomo «acronimo» che deve agire da funzione, ma il suo *valore*, sia esso «macaroni», «cheese» o quello che volete.

Ricorriamo a un piccolo trucco: se il valore dell'atomo «acronym» è «rhubarb» e se scrivo «(list acronimo)», il valore che il genietto mi darà sarà la lista «(rhubarb)». Il genietto, tuttavia, lo considererà semplicemente un inerte pezzo del Lisp e non un comando che desideravo eseguisse. Non può leggermi nel pensiero; come ottenere che compia l'operazione desiderata? La soluzione sta nel ricordarsi dell'esistenza della funzione «eval», che porta il genietto a considerare una struttura di dati come un ordine da eseguire. In questo caso devo semplicemente dire, «(eval (list acronimo))» per ottenere la lista «(ravioli, heated under butter and rhubarb (basta!))». E se l'acronimo avesse avuto un valore differente, il genietto mi avrebbe passato una lista differente.

Ora sappiamo abbastanza per costruire una funzione che sviluppi gli acronimi reciprocamente ricorsivi in espressioni lunghe, ma finite, la cui grandezza e struttura sono determinate da molti «lanci» della moneta «rand». Non vi condurrò passo per passo nella costruzione di questa funzione, ma ve la mostrerò soltanto e vi lascerò il compito di analizzarla. È modellata molto fedelmente sulla precedente funzione «rimpiazza».

```
(def sviluppa
 (lambda (espressione probabilità
  (cond
   ((atomo espressione) espressione)
   ((è-un-acronimo (car espressione))
    (cond
     ((lessp (rand) probabilità)
      (append
       (sviluppa (eval (list (car espressio-
ne)))
                 (minore probabilità))
       (sviluppa (cdr espressione) proba-
bilità)))
      (t
       (con (car espressione)
            (sviluppa (cdr espressione)
                      probabilità)))))
   (t
    (cons
     (sviluppa (car espressione)
               (minore probabilità))
     (sviluppa (cdr espressione) probabi-
lità))))
))
```

Si noti che «sviluppa» ha due parametri: uno rappresenta l'espressione da sviluppare, l'altro rappresenta la probabilità di sviluppo degli acronimi che appartengono ai livelli alti dell'espressione data. (Allora il valore dell'atomo «probabilità» sarà sempre un numero reale compreso tra 0 e 1). Come avviene per la sequoia, lo sviluppo probabilmente andrà diminuendo man mano che le chiamate diventano sempre più ricorsive. Ecco perché la richiesta di sviluppo di (car espressione) ha una probabilità *minore*. Per essere precisi possiamo definire la funzione «minore» nel modo seguente:

```
(def minore (lambda (x) (times x .8)))
```

Ogni volta che un acronimo si sviluppa, la sua progenie ha una probabilità di sviluppo che è solo 0,8 volte quella dell'acronimo genitore. Ciò significa che gli acronimi annidati più internamente hanno una probabilità molto bassa di svilupparsi ulteriormente. Potreste usare un fattore di riduzione qualsiasi: nulla impone il fattore 0,8, se non il fatto che mi ha dato buoni risultati.

Nella definizione precedente rimane una sola funzione da descrivere: è «è-un-acronimo». Il nome si spiega da solo: per prima cosa la funzione verifica se il suo argomento è un atomo; se non lo è, dà come valore nil. Se l'argomento *è* un atomo la funzione procede a vedere se l'atomo ha una definizione di funzione, in particolare, una definizione a forma di acronimo. Se la risposta è positiva, «è-un-acronimo» dà valore t; altrimenti dà nil. Come avvenga precisamente la cosa dipende dalla vostra variante del Lisp; ecco perché non ve l'ho mostrato esplicitamente. Nel Franz Lisp basta una riga.

Forse avete notato che vi sono due clausole condizionali molto vicine che iniziano con «t». Come è possibile che ci siano due «otherwise», uno dietro l'altro? In realtà essi appartengono a due «cond» diversi, uno annidato nell'altro. Il primo «t» (che appartiene al «cond» più interno) si applica a un caso in cui abbiamo a che fare con un acronimo, ma nel quale la nostra moneta dà croce (il che significa che non si deve sviluppare); il secondo «t» (che appartiene al «cond» più esterno) si applica a un caso in cui si sia scoperto di non avere affatto a che fare con un acronimo.

Se l'esaminate con cura, «sviluppa» ha una sua logica interna perfettamente sensata. È, tuttavia, anche vero che, per quanto lo esaminiate con cura, il risultato prodotto da «sviluppa», usando questa famiglia di acronimi, resta abbastanza folle. Eccovi un esempio:

(rich italian green and red linguini and shad and unusual choice of fine flavors, particularly excellent, strong, powerful, rich espresso, suppressing sleep outrageously (eccellente!) and green and red lasagne in cheese (noodles everywhere!) in cheddar, havarti, Emmentaler (particularly sharp Emmentaler) noodles (oodles of delicious linguini), elegantly served (oodles of delicious linguini), elegantly served (oodles of delicious linquini and sauce and garlic (noodles (oodles of delicious linguini), elegantly served everywhere!) and meatballs, casually heaped onto pasta and sauce (that's all!) and sauce (that's a luscious lunch) sauce (including noodles (oodles of delicious linguini), elegantly served), got usually in northern Italy), elegantly served over crisp cheese, heated immediately and tomatoes on macaroni and cheese (a repast of Naples, Italy) (and tomatoes only), exquisitely sweetly pickled in cheese endive dressing (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and rich Italian gnocchi and tomatoes (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and rigatoni and vongole in oil, lavishly introduced, heated under butter and ravioli, heated under butter and rich Italian garlic noodles over crisp cheese, heated immediately and tomatoes (or noodles instead) and vongole in oil, lavishly introduced, heated under butter and ravioli, heated under butter and rhubarb (basta!) (basta!) (basta!) (basta!) (belly all stuffed (tummy ache!)) (basta!))

Riuscite a stabilire da quale acronimo sia nata questa mostruosità gastronomica? Come potete vedere, il Lisp non è esattamente il linguaggio di programmazione da scegliere se si vuole perdere peso. Qui di seguito, ecco un ultimo esempio dei fasti degli spaghetti ricorsivi, esito dello sviluppo di un diverso punto di partenza:

(macaroni and cheese (a rather extraordinary pasta and sauce, typical of Naples, Italy) and cheddar, havarti, Emmentaler (particularly sharp Emmentaler) (a rather extraordinary pasta and shad and unusual coffee (eccellente!) (that's luscious lunch) and sauce (that's all!) and shad and unusual choice of fine flavors, particularly espresso (eccellente!) (that's all!) and sauce (that's a luscious lunch) and spaghetti, heated al dente and unusual choice of fine flavors, particularly excellent, strong, powerful, rich espresso, suppressing sleep outrageously (eccellente!), typical of Naples, Italy))

La funzione «sviluppa» mette in luce una delle caratteristiche più potenti del Lisp, vale a dire l'abilità che un programma del Lisp ha di prendere le strutture di dati che ha creato e trattarle come pezzi di codice (cioè passarle al genietto del Lisp come comandi). Qui la cosa è stata fatta in maniera molto rudimentale. Un atomo è stato chiuso tra parentesi e la minuscola lista risultante è stata valutata. Per costruire la struttura dei dati non è stato necessario, in questo caso, molto lavoro, ma altre volte si possono consegnare al genietto del Lisp, perché li valuti, elaborati pezzi di struttura costruiti applicando ripetutamente la funzione «cons». Tali pezzi di codice potrebbero essere definizioni di nuove funzioni o tante altre cose. L'idea di fondo è che nel Lisp si ha la possibilità di elevare un'inerte struttura di dati, che contengono informazioni, al livello di un «agente animato» che diventa a sua volta manipolatore di strutture inerti. Questa circolarità, questo anello programma-dati può continuare in eterno con strutture che si allungano, si ripiegano indietro, modificando indirettamente se stesse o strutture collegate.

Alcuni tipi di strutture di dati inerti o passive contenenti informazioni vengono talora chiamate «conoscenza dichiarativa»; «conoscenza» perché codificano fatti, relativi al mondo, a cui si può accedere cercandoli in un indice alla maniera in cui i fatti «imparati sui libri» sono accessibili a un essere umano, «dichiarativa» perché hanno spesso una forma che assomiglia a quella di un enunciato dichiarativo. I pezzi di codice animati o attivi, invece, vengono detti «conoscenza procedurale»; «conoscenza», perché incorporano l'insieme di abilità del programma, un po' come le abilità inconsce di un essere umano, acquisite un tempo attraverso una lunga serie di esercitazioni meccaniche, «procedurale» perché definiscono sequenze di azioni («procedure») che manipolano effettivamente le strutture di dati. Talora questi diversi tipi di conoscenza vengono detti «conoscenza del che cosa» e «conoscenza del come».

Questa distinzione ricorderà ai biologi quella tra i geni che sono strutture interne alla cellula relativamente inerti e gli enzimi che sono tutt'altro che inerti. Gli enzimi sono gli agenti animati della cellula: manipolano e trasformano, in modi incredibilmente raffinati, tutte le strutture inerti. Inoltre l'anello del Lisp programma-dati dovrebbe ricordare ai biologi il modo in cui i geni determinano la forma degli enzimi e gli enzimi manipolano i geni (tra le altre cose). Allora l'anello procedurale-dichiarativo, programma-dati nel Lisp rappresenta un esempio primitivo, ma molto concreto e utile di una delle configurazioni fondamentali che stanno alla base della vita: la capacità che hanno le strutture passive di controllare il proprio destino, creando e regolando strutture attive, di cui esse stesse determinano la forma.

Abbiamo parlato sempre del genietto del Lisp come di un misterioso agente dato, senza chiederci da dove venga e che cosa lo faccia funzionare. Una delle proprietà più interessanti del Lisp è che si può descrivere pienamente e con facilità la natura del genietto del Lisp all'interno del Lisp stesso. In altre parole, l'interprete del Lisp si può facilmente descrivere in Lisp. Ovviamente, se non ci fosse nessun interprete del Lisp che interpretasse *quel l'*interprete del Lisp, la cosa potrebbe sembrare un esercizio assurdo e inutile, un po' come descrivere agli stranieri in un inglese fiorito il miglior modo per imparare l'inglese. La cosa non è però così stupida come potrebbe sembrare.

In primo luogo, se conoscete un po' di inglese, potete avanzare ulteriormente nella vostra conoscenza; c'è un punto oltre il quale le spiegazioni scritte in inglese sono molto utili e, ciò che più conta, questo

punto non è molto lontano dal livello iniziale. Pertanto, dovete in precedenza acquisire, e potete farlo autonomamente, soltanto il «nucleo» e poi potete procedere da soli. Per i bambini, quando incominciano a leggere, è eccitante imparare da soli nuove frasi semplicemente per averle incontrate parecchie volte di seguito. Il loro vocabolario si accresce con straordinaria rapidità. La stessa cosa avviene quando c'è il nucleo del Lisp in un sistema; la rimanente interpretazione del Lisp si può scrivere in Lisp e generalmente avviene proprio così.

La ragione per cui è facile scrivere l'interprete del Lisp nel Lisp stesso sta nel suo anello programma-dati. Il Lisp si presta per la stesura di interpreti per tutti i tipi di linguaggi dei calcolatori. Ciò significa che può servire come base su cui costruire altri linguaggi.

Per esprimere la cosa in maniera più concreta, supponete di aver preparato sulla carta un nuovo linguaggio, il «Flumsy». Se sapete veramente come funziona, non dovrebbe riuscirvi troppo difficile scriverne un interprete in Lisp. Una volta completato il vostro interprete, nascerà un nuovo genietto al quale potrete esprimere i vostri desideri in Flumsy e che li comunicherà al genietto del Lisp in Lisp. Ovviamente tutti i meccanismi che permettono al «metagenietto» del Flumsy di parlare con il genietto del Lisp sono essi stessi elaborati dal genietto del Lisp. È una pura finzione? Parlare in Flumsy è solo un modo di parlare in un Lisp camuffato?

Quando, alle trattative per gli armamenti, gli americani parlano con le loro controparti russe servendosi di un'interprete, parlano davvero un russo camuffato? Oppure il nocciolo della questione sta nel chiedersi se la lingua madre dell'interprete sia l'inglese o il russo, su quale dei due l'altra lingua è stata costruita come seconda lingua? E supponete di scoprire che in realtà la lingua madre dell'interprete sia il lituano, che abbia imparato l'inglese solo da ragazza e che abbia poi imparato il russo nelle scuole superiori, in cui tale insegnamento le veniva impartito in inglese; penserete allora che quando l'interprete parla russo, parli in un inglese camuffato, o, peggio, che parli un lituano doppiamente camuffato?

Analogamente, potreste trovare che l'interprete del Lisp è scritto in realtà in Pascal o in qualche altro linguaggio di calcolatore. E poi qualcuno potrebbe far cadere anche la maschera del Pascal e rivelarvi che tutte le istruzioni sono in realtà scritte in *linguaggio di macchina* e che quindi vi state completamente sbagliando se pensate che il calcolatore stia parlando in Flumsy, in Pascal o in qualsiasi altro linguaggio di alto livello.

Quando un interprete gira su un altro, c'e sempre il problema di dove fermarsi nel risolvere i vari livelli. Per quanto mi riguarda, raramente penso a che cosa sta sotto l'interprete del Lisp e, quando lavoro con il sistema del Lisp, mi comporto come se stessi parlando con «qualcuno» la cui «lingua madre» sia il Lisp, proprio come, quando tratto con una persona, raramente penso a ciò di cui è fatto il suo cervello e non la riduco mentalmente a un mucchio di strutture neuronali in agitazione. È naturale per il mio sistema percettivo riconoscerla a un certo livello e non scavare al di sotto di esso.

Se qualcuno scrivesse un programma in grado di trattare in cinese semplici domande e risposte riguardanti i ristoranti da visitare e se quel programma fosse scritto in un ipotetico linguaggio SEARLE (per «Simulated East-Asian Restaurant-Lingo Expert», esperto di linguaggi culinari dell'Asia orientale simulato), potrei scegliere di considerare il sistema come un parlante nativo cinese (presupponendo che possa fornire una prestazione credibile e non troppo lunga) o come un vero parlante SEARLE. Posso cambiare punto di vista a mio piacimento. La mia scelta è regolata per lo più da fattori pragmatici, come il tema che mi interessa di più al momento (i ristoranti cinesi o il funzionamento della grammatica), in che misura la velocità del livello dato corrisponde a quella del mio cervello e, non ultimo, se parlo meglio il cinese o il SEARLE. Se per me il cinese è un insieme di suoni senza senso, preferirò il SEARLE; se, invece, il SEARLE per me è un insieme di oscure espressioni tecniche, sceglierò il cinese. E se scopro che il sistema SEARLE è, a sua volta, costruito a partire dal Lisp, avrò un terzo punto di vista da scegliere, e così via.

Ammassando interpreti su interpreti, però, si perde rapidamente efficienza. È come far funzionare un motore ricavando energia da una serie di generatori elettrici, ciascuno dei quali alimentato dal precedente: a ogni passaggio si perde un bel po' di energia. Per i generatori di solito non c'è bisogno di un lungo ricorso all'indietro, ma per gli interpreti è spesso l'unica via. Se non c'è alcun calcolatore che abbia come proprio linguaggio il Lisp, costruite un interprete del Lisp per il calcolatore che avete a disposizione e fate girare il Lisp in tal modo. Se volete il Flumsy o il SEARLE, poi, potete costruirli al di sopra di questa «Lisp machine», virtuale. Un procedimento così indiretto può essere pesantemente inefficiente e può rendere la vostra nuova «Flumsy machine» o «SEARLE machine» decine di volte più lenta di quel che vorreste.

Negli ultimi anni si sono verificati importanti sviluppi nella tecnologia dei calcolatori e oggi esistono calcolatori basati sul Lisp a livello hardware. Ciò significa che questi calcolatori «parlano» il Lisp in un senso più profondo (potremmo dire «più correntemente») delle Lisp machine virtuali. Significa anche che, quando lavorate su un calcolatore del genere, siete «immersi» in un ambiente Lisp. Un ambiente Lisp va notevolmente oltre ciò che ho descritto fin qui, perché è più che un linguaggio per scrivere programmi. Comprende un programma di «editing» con il quale si può creare e modificare il proprio programma (e dei testi), un programma di revisione con il quale si possono facilmente localizzare ed eliminare gli errori e molte altre caratteristiche tutte reciprocamente compatibili e compatibili con una sovrastante «filosofia del Lisp».

Calcolatori cosiffatti, benché ancora costosi e in un certo senso sperimentali, stanno diventando sempre più economici e disponibili. Sono prodotti da varie nuove aziende come la LMI (Lisp Machine Inc.), la Symbolics Inc., entrambe di Cambridge, Massachusetts e da aziende tradizionali come la Xerox. Il Lispsi può anche trovare sulla maggior parte dei calcolatori personali.

Perché, in conclusione, il Lisp è così diffuso nel campo dell'intelligenza artificiale? Non c'è un'unica risposta. Sarebbe giusto dire che il Lisp è «il linguaggio del pensiero»? Certamente no. Gli studiosi di intelligenza artificiale avrebbero potuto esprimersi così 10 o 20 anni fa, ma oggi pochi sarebbero così tronfi. Una ragione per cui il Lisp si è mantenuto così popolare è che, come ho scritto nel mio primo articolo sull'argomento, il Lisp è chiaro. Inoltre il Lisp può essere così elegante che una buona funzione del Lisp può dare a un appassionato lo stesso piacere estetico che una lirica dà a un amante della poesia. Un mio collega, Dan Friedman, sta scrivendo, infatti, un libro intitolato *Lisp Poems* e invita i lettori interessati a inviargli i loro risultati più eleganti. Il suo indirizzo è: Computer Science Department, Indiana University, Bloomington, Ind. 47405.

Quali sono allora le proprietà del Lisp, che lo hanno reso centrale per le ricerche sull'intelligenza artificiale? Direi che quelle indicate qui di seguito sono più che sufficienti:

1. il Lisp è chiaro;
2. il Lisp è interattivo;
3. il Lisp è incentrato sull'idea delle liste e sulla loro manipolazione e le liste sono strutture di dati estremamente flessibili;
4. il codice del Lisp, avendo la stessa forma dei dati del Lisp, può essere facilmente preparato e fatto girare in Lisp;
5. si possono facilmente costruire e sperimentare in Lisp interpreti per nuovi linguaggi;
6. l'«immersione» in un ambiente del tipo Lisp risulta naturale a molte persone;
7. il Lisp è permeato dallo «spirito della ricorsività».

Forse proprio questa affermazione un po' intangibile coglie meglio il senso del Lisp. Per qualche motivo molti ricercatori nel campo dell'intelligenza artificiale sembrano avere profonda sensazione che la ricorsività, in una forma o nell'altra, sia connessa all'«astuzia» dell'intelligenza. È un'intuizione, una sensazione vaga e un po' mistica, che io certamente condivido; ma è da vedere se si rivelerà fruttuosa sul lungo periodo.

[P.S. Nell'articolo di Hofstadter del mese scorso, a pagina 119, si parlava genericamente di un'istruzione «traccia». Poiché l'esposizione poteva essere ambigua, ricordiamo che il Lisp possiede una funzione standard che si chiama in effetti «trace». *n.d.r.*]

suo terreno di esercitazioni, preoccupandosi della propria immagine molto di più che della sorte delle stupide pedine.
(C. Bernardini)

MATERIE PRIME ED ECONOMIA MONDIALE

di Enzo Grilli
Il Mulino, Bologna, 1982, pp. 394 (L. 25 000).

Enzo Grilli è conosciuto per aver diretto l'Ufficio studi della Confindustria e per essere stato recentemente nominato segretario generale per la programmazione economica presso il Ministero del bilancio. Ma è soprattutto la sua esperienza alla Banca Mondiale come direttore della Divisione materie prime e politiche commerciali che caratterizza l'impostazione del suo libro sulle materie prime. Utilizzando un punto di osservazione come la Banca Mondiale, il tema viene infatti affrontato con un'angolazione più internazionale di quanto non avvenga in altri studi sull'argomento.

Una tale impostazione presenta vantaggi e svantaggi. Da un lato consente all'autore di non perdere di vista l'insieme dei molteplici aspetti dell'economia delle materie prime e di spaziare dall'esame delle produzioni e del commercio internazionale al tema dei rapporti tra produzioni primarie, commercio estero e sviluppo, all'analisi della struttura dei mercati, alla discussione sui problemi di sicurezza degli approvvigionamenti, sul dialogo Nord-Sud e sulla cooperazione internazionale; consente anche di procedere a una rassegna interessante, anche se non esaustiva, di alcuni dibattiti di carattere generale ancora aperti, come quelli relativi allo sviluppo complesso, all'andamento delle ragioni di scambio, ad alcuni aspetti della teoria dei vantaggi comparati nel commercio internazionale. D'altra parte la medesima impostazione rischia di accentuare la tendenza a mediare posizioni teoriche e politiche, fino a rendere asettiche conclusioni e proposte.

Il libro si caratterizza per il fatto di considerare insieme tutte le materie prime, comprendendo quindi i prodotti alimentari e quelli agricoli di uso industriale, i prodotti minerari energetici e non energetici. Alcuni tipi di analisi di carattere generale possono certamente effettuarsi considerando le materie prime nel loro complesso; e anzi, in tal modo, possono essere meglio articolate definizioni come quella di «cartellizzabilità» o di elasticità della domanda e dell'offerta, si può più correttamente discutere aspetti comuni alle diverse economie sottosviluppate e relazioni tra settore primario e sviluppo. Ma il non considerare la specificità dei vari gruppi di prodotti primari per quanto concerne le caratteristiche tecniche, dell'offerta, della domanda e delle utilizzazioni finali, oltre che per le differenti politiche configurabili, rende problematico l'approfondimento dello studio della struttura dei mercati e trascura le difformità importanti, all'interno del sottosviluppo, tra paesi agricoli, minerari e produttori di petrolio. Le conseguenze di questo tipo di aggregazione sono mitigate dal fatto che i principali prodotti agricoli e alcuni metalli di più largo uso industriale vengono brevemente considerati singolarmente.

Guardando in particolare ai prodotti minerari non energetici, l'esposizione di Grilli risulta chiara e sintetica, anche se si avverte l'esigenza di qualche riferimento - anche a prescindere dalle infrastrutture e dalle sostituzioni di alcuni materiali e dal riciclo - alle relazioni con altri settori come quello energetico, i trasporti, i settori industriali utilizzatori.

L'analisi dei prezzi è probabilmente l'aspetto più approfondito e interessante del libro. Oltre a toccare il tema dell'inflazione importata, l'autore studia le elasticità della domanda e dell'offerta rispetto al reddito e ai prezzi, e individua, per quanto riguarda i prezzi di un insieme di minerali e metalli, sia le tendenze di lungo e di medio periodo, sia l'andamento ciclico, indicandone le principali variabili esplicative sulla base di stime econometriche e giovandosi dei suoi precedenti lavori sull'argomento. Tuttavia, il non aver distinto i dati relativi a minerali e a metalli non soltanto limita l'analisi dei mercati e la validità delle relazioni funzionali stimate, ma rende non univoche anche osservazioni di carattere generale come, ad esempio, sulla ripartizione di produzioni e scambi internazionali tra paesi in via di sviluppo e paesi industrializzati o sul grado di protezione tariffaria (nominale ed effettiva) dei paesi industrializzati. Non distinguendo tra attività mineraria, metallurgica e di prima trasformazione viene inoltre ignorata la questione fondamentale del *processing capacity gap* dei paesi in via di sviluppo a economia mineraria.

Quest'ultima circostanza è abbastanza sorprendente se si considera quanto l'autore si sforzi nell'elencare e spiegare, mantenendosi sopra le parti, le contrapposte argomentazioni di politica per le materie prime del dibattito Nord-Sud. In realtà l'esame di tali contrapposizioni - che occupa l'ultima parte del libro - si limita sostanzialmente a riprendere i termini della discussione nell'ambito degli accordi di stabilizzazione relativi a singoli prodotti (stagno, gomma ecc.) e dei negoziati UNCTAD per il Programma comune per le materie prime, di cui peraltro vengono descritte efficacemente le fasi salienti.

Se è vero che il trasferimento nei paesi in via di sviluppo di capacità di raffinazione, *processing* e prima trasformazione industriale di prodotti minerari non è neppure entrato nelle agende dei negoziati multilaterali, è pure evidente trattarsi di una questione che è cruciale per i paesi emergenti a economia mineraria e che dai più attenti paesi industrializzati - il Giappone è un caso istruttivo - viene affrontata concretamente con politiche nazionali di settore e con accordi bilaterali con i paesi in via di sviluppo.

Da un libro che anche nel titolo sottolinea il ruolo delle materie prime nell'economia mondiale ci si attendeva forse che, nell'individuare le possibili aree di conflitto e di cooperazione, venissero evidenziate anche le politiche nazionali e le strategie d'investimento. L'autore ha preferito concentrarsi nell'esame dell'esperienza della cooperazione multilaterale: da qui anche il suo distaccato pessimismo. *(G. Mureddu)*

IL SOLE

di Giovanni Godoli
Einaudi, Torino, 1982, pp. 284
(L. 12 000).

Per apprezzare compiutamente certe opere di divulgazione scientifica degli ultimi anni può essere utile sfogliare alcuni testi analoghi apparsi nel periodo immediatamente precedente e seguente l'ultima guerra mondiale. Rileggere oggi la divulgazione dei grandi scienziati della prima metà del secolo, da Planck a Jeans, da Gold a Einstein stesso serve a evidenziare gli spunti di innovazione più significativi in alcuni libri scritti dai loro successori proprio in questi anni ottanta. Al primo sguardo le difficoltà sembrano aumentate: si trattano argomenti che prima venivano accantonati perché inaccessibili al lettore, si preferiscono immagini più approfondite e articolate alle metafore intuitive. In parte questa è naturalmente un'illusione, perché molti problemi oggi abbastanza noti al grosso pubblico oltre che agli specialisti si trovavano trenta o quarant'anni fa alle frontiere della ricerca. Eppure esistono anche differenze di impostazione reali, e la più evidente è davvero una specie di rinuncia alla facilità: la semplicità concettuale della scienza cui Einstein teneva tanto passa in secondo piano, al lettore si richiede anche l'impegno e la disponibilità alla fatica. D'altro canto, però, non si offre solo la comprensione di un'idea («che cosa vuol dire» la relatività), ma si cerca di spiegare come funziona, come costruisce le sue conclusioni una particolare disciplina. Forse questo accade perché lo scienziato avverte l'esigenza di farsi capire e giudicare dal cittadino, in un momento di forti perplessità nei confronti di certe attività di ricerca: certo il pubblico sembra premiare questo sforzo, e diversi volumi di divulgazione scientifica sollevano l'attenzione dei quotidiani e dei settimanali, riscuotono un successo editoriale magari imprevisto.

È anche il caso di questo libro sul Sole, un tema non ancora affrontato dalla letteratura divulgativa come problema complesso e autonomo. L'autore è uno dei direttori dell'Osservatorio solare europeo, un grande progetto a carattere sovranazionale, e titolare dell'unica cattedra in Europa di fisica solare; qui intende sfruttare un momento di stasi nei programmi di osservazione spaziale per una panoramica sull'argomento. Il risultato, per chi si avvicina per la prima volta all'astronomia, è un'esposizione aggiornata, valida probabilmente fin verso gli anni novanta, mentre chi già se ne interessa, per studio o per divertimento, ha a disposizione un completo e prezioso riepilogo delle conclusioni più recenti. La puntualizzazione del ruolo della fisica solare nella cultura prima che nell'astronomia apre una trattazione organizzata in modo originale, profondamente valido e stimolante: la struttura del testo mira a facilitare l'orientamento del lettore all'interno della disciplina, lasciandolo il più possibile libero di seguire i propri interessi. Nella sezione introduttiva, infatti, sono fornite le nozioni necessarie per approfondire anche singolarmente i capitoli successivi, ciascuno dei quali affronta un problema autonomo; questa prima parte, che comprende anche un breve sommario dei principali dati sul Sole, è imperniata però su una descrizione dei metodi di indagine propri della fisica solare e dell'astronomia in genere. In questo modo il fattore strumentale assume esplicitamente un ruolo cruciale: risulta chiaro fin dall'inizio il grado di attendibilità di ogni affermazione, come pure la distinzione tra ipotesi avanzate a titolo di tentativo e interpretazioni generalmente accettate di fenomeni abbastanza ben noti. Le nozioni generali di fisica e di astronomia sono distribuite in note ampie ed esaurienti e permettono, a chiunque sia disposto ancora a profondere un certo impegno, di colmare le eventuali lacune.

Il corpo principale dell'opera si articola in tre parti: interno del Sole; atmosfera solare; regioni solari attive. Nella prima, oltre a descrivere la struttura interna del Sole e la sua probabile evoluzione, si parla dei campi magnetici, della rotazione e del conseguente schiacciamento polare. Un posto di rilievo è riservato al problema dell'emissione di neutrini, un problema irrisolto e perciò affrontato frettolosamente su testi anche più avanzati. La descrizione dell'atmosfera solare prende in esame separatamente la fotosfera, la cromosfera e la corona, soffermandosi sulla struttura interna e sui modelli attualmente disponibili di ognuna. Anche in questo caso un problema aperto come quello della temperatura della corona viene affrontato come tale e le soluzioni proposte sono discusse senza troppe illusioni. Forse molti preferiranno però leggere per primi i capitoli sulle regioni attive, nei quali vengono descritti i fenomeni resi celebri ormai da decenni dai progressi della fotografia astronomica: le macchie, le facole, le protuberanze e i brillamenti. Il testo indugia a lungo sulle procedure con cui si tenta di valutare quantitativamente l'intensità dell'attività solare in ciascuna forma. L'analisi dei concetti di area proiettata e area corretta o dei sistemi di classificazione dei brillamenti, ad esempio, offre un'immagine viva dei rischi di soggettività ancora legati all'osservazione astronomica, introducendo criticamente il problema dei cicli di attività solare e del loro studio.

Gli argomenti più nuovi e stimolanti si incontrano negli ultimi capitoli. La fisica Sole-Terra, con le sue applicazioni anche tecnologiche immediate, è un campo di ricerca relativamente nuovo, presentato forse per la prima volta in modo così completo: questa particolare branca della fisica è un tipico esempio di quelle attività interdisciplinari che oggi sono tra le più feconde di tutta la scienza. La ricerca di fenomeni analoghi a quelli dell'attività solare su altre stelle, d'altro canto, oltre a essere una sfida ai limiti estremi delle possibilità tecniche degli strumenti, e delle capacità umane di interpretarne il responso, offre un'occasione di riflessione riportandoci a una prospettiva più generale: capire come apparirebbero i fenomeni solari da distanze dell'ordine di qualche anno luce è un esercizio di astrazione e di obiettività sia per lo specialista sia per l'uomo di cultura.

Il Sole è anche un volume particolarmente curato dal punto di vista editoriale, tenendo specificamente in considerazione le sue finalità divulgative: il corredo iconografico è esauriente e stimolante; nelle successive edizioni sarà molto utile anche un indice analitico che permetta di destreggiarsi nella ricchezza delle informazioni. In un'opera di divulgazione valida e nuova come questa, però, assume particolare rilievo il problema della leggibilità, che interessa tutti i libri scientifici scritti in italiano per i non specialisti. Da un lato infatti, come si è visto, è giusto rifiutare il semplicismo e adottare un linguaggio anche complesso, che permetta di maneggiare agevolmente concetti elaborati; d'altra parte l'italiano scientifico è una lingua difficile, per diversi motivi legati al fatto che lo scienziato attivo in Italia scrive e pensa in inglese. Questo significa innanzitutto che i vocaboli più appropriati, quelli che tengono meglio conto degli sviluppi teorici più recenti, sono costruiti secondo regole di formazione lontane dalla sensibilità della nostra lingua: si pensi a termini come «massivo» o proprio a «fisica Sole-Terra». La struttura del discorso, poi, è appesantita dalla necessità di sviluppare come proposizioni subordinate i sostantivi anglosassoni con funzione attributiva, e dalla tendenza a coordinare in periodi articolati le brevi frasi tipiche dello stile americano. Tutto procede così in modo perfettamente funzionale per chi vi sia abituato, ma profondamente estraneo al lettore generico, costretto a faticose interruzioni e riletture. Il problema si fa ancora più grave quando si traduce in italiano un testo inglese: la scelta più naturale è allora la lingua sintatticamente ineccepibile, ma a volte perfino sgradevole in cui i temi scientifici sono effettivamente dibattuti. Un'impresa culturale come è la divulgazione scientifica moderna deve raccogliere questa sfida. Perché esista una cultura scientifica italiana occorre anche una lingua adatta: alla sua elaborazione contribuiscono le redazioni dei periodici e delle case editrici scientifiche, ma è indispensabile che anche lo scienziato si sforzi di dare il proprio contributo. *(G. Marcellino)*